Insights into the Effects of Long-Term Artificial Selection on Seed Size in Maize

Candice N. Hirsch,* Sherry A. Flint-Garcia,†,‡ Timothy M. Beissinger,§, Steven R. Eichten,†† Shweta Deshpande,‡‡ Kerrie Barry,‡‡ Michael D. McMullen,†,‡ James B. Holland,†,§§ Edward S. Buckler,†,***,††† Nathan Springer,†† C. Robin Buell,‡‡‡,§§§ Natalia de Leon,§,**** and Shawn M. Kaeppler§,****,1**

*Department of Agronomy and Plant Genetics and ††Department of Plant Biology, University of Minnesota, Saint Paul, Minnesota 55108, †United States Department of Agriculture, Agricultural Research Service, Columbia, Missouri 65211, ‡Division of Plant Sciences, University of Missouri, Columbia, Missouri 65211, §Department of Agronomy, **Department of Animal Sciences, and ****Department of Energy Great Lakes Bioenergy Research Center, University of Wisconsin, Madison, Wisconsin 53706, ‡‡Department of Energy, Joint Genome Institute, Walnut Creek, California 94598, §§Department of Crop Science, North Carolina State University, Raleigh, North Carolina 27695, ***Institute for Genomic Diversity and †††Department of Plant Breeding and Genetics, Cornell University, Ithaca, New York 14853, and ‡‡‡Department of Plant Biology and §§§Department of Energy Great Lakes Bioenergy Research Center, Michigan State University, East Lansing, Michigan 48824

ABSTRACT Grain produced from cereal crops is a primary source of human food and animal feed worldwide. To understand the genetic basis of seed-size variation, a grain yield component, we conducted a genome-wide scan to detect evidence of selection in the maize Krug Yellow Dent long-term divergent seed-size selection experiment. Previous studies have documented significant phenotypic divergence between the populations. Allele frequency estimates for ~3 million single nucleotide polymorphisms (SNPs) in the base population and selected populations were estimated from pooled whole-genome resequencing of 48 individuals per population. Using F_{ST} values across sliding windows, 94 divergent regions with a median of six genes per region were identified. Additionally, 2729 SNPs that reached fixation in both selected populations with opposing fixed alleles were identified, many of which clustered in two regions of the genome. Copy-number variation was highly prevalent between the selected populations, with 532 total regions identified on the basis of read-depth variation and comparative genome hybridization. Regions important for seed weight in natural variation were identified in the maize nested association mapping population. However, the number of regions that overlapped with the long-term selection experiment did not exceed that expected by chance, possibly indicating unique sources of variation between the two populations. The results of this study provide insights into the genetic elements underlying seed-size variation in maize and could also have applications for other cereal crops.

GRAIN produced by cereal crops is a staple food source in many regions of the world in terms of direct human consumption and as an animal feed source. Understanding the molecular mechanisms underlying cereal grain yield and exploiting that knowledge through improved cultivars is essential to providing a stable food source to an ever-growing human population. Yield-component traits are of particular interest, as they generally have a higher heritability than grain yield *per se* (Austin and Lee 1998). For example, increasing seed size has been hypothesized as one method for increasing grain yield in cereal crops (Odhiambo and Compton 1987; Kesavan *et al.* 2013), and positive correlations between seed size and grain yield have been shown in maize (Peng *et al.* 2011) as well as other cereals such as *Sorghum bicolor* (L.) Moench (Yang *et al.* 2010). Maize is a prime species with which to explore natural and artificial variation related to grain-yield and yield-component traits in the cereals, as it is the most widely grown cereal crop worldwide and has vast genetic resources for probing the genetic basis of seed traits.

doi: 10.1534/genetics.114.167155
Manuscript received June 12, 2014; accepted for publication July 8, 2014; published Early Online July 17, 2014.
Supporting information is available online at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.167155/-/DC1.
Sequence data from this article have been deposited with the Sequence Read Archive at the National Center for Biotechnology Information study under accession no. SRP013705.
[1]Corresponding author: Department of Agronomy, 1575 Linden Dr., University of Wisconsin, Madison, WI 53706. E-mail: smkaeppl@wisc.edu

The maize seed is composed of the embryo and endosperm that develop from double fertilization, the aleurone, which is an epidermal layer that covers the endosperm, and the maternal pericarp tissue. The endosperm, the primary storage component of the seed in maize, consists primarily of starch, while the embryo is high in oil content (Kiesselbach 1999). Storage proteins also accumulate in the developing endosperm of maize, with the main class of storage proteins being zeins (Paulis and Wall 1977). Large effect mutants such as *Miniature1* (*Mn1*) (Cheng *et al.* 1996), *opaque-2* (*o2*) (Schmidt *et al.* 1990), *shrunken-2* (*sh2*) (Bhave *et al.* 1990), *stunter1* (*stt1*) (Phillips and Evans 2011), *Zea mays Outer Cell Layer1* (*ZmOCL1*) (Khaled *et al.* 2005), and others (Neuffer *et al.* 1997) have been identified and affect overall seed and/or endosperm development in maize. Additionally, recent work has begun to elucidate the regulatory networks involved in maize seed development (Fu *et al.* 2013). Despite these studies on overall seed development, the genetic basis of seed-size variation in maize and other cereal crops is still largely unknown.

Selection increases the frequency of favorable alleles in a population. Therefore, the assessment of allele frequency change is a useful technique for identifying genomic regions that were targeted by selection (Lewontin 1962). Specific methods vary depending on the populations under study and the genotyping methods employed (Wright 1951; Akey *et al.* 2002; Sabeti *et al.* 2002; Oleksyk *et al.* 2008; Wisser *et al.* 2008; Turner *et al.* 2011). For example, in natural populations, statistics that measure population divergence such as F_{ST} (Wright 1951) can be calculated and loci displaying extreme values above an empirically determined genome-wide threshold are implicated as potentially associated with selection (Akey *et al.* 2002; Oleksyk *et al.* 2008). Identification of selection signatures has successfully been used to reveal the genetic basis of several traits across numerous species, including heat tolerance in yeast (Parts *et al.* 2011), body-size variation in *Drosophila melanogaster* (Turner *et al.* 2011) and chickens (Johansson *et al.* 2010), milk production in Holstein cattle (Pan *et al.* 2013), and prolificacy (Beissinger *et al.* 2014) and northern leaf blight resistance (Wisser *et al.* 2008) in maize.

The goal of this study is to dissect the genetic architecture of seed-size variation in cereal crops using maize as a model. Long-term artificial-selection experiments contain a wealth of information about trait architecture and, with the advent of next-generation sequencing, we can now harness that information. To unravel the genetic architecture of seed-size variation in maize, we compared pooled whole-genome resequencing data from populations from a divergent selection experiment for small and large seed size (Odhiambo and Compton 1987; Russell 2006) (Figure 1). Previous work has demonstrated significant phenotypic variation among the three Krug populations for seed weight and other morphological and compositional traits (Sekhon *et al.* 2014). In this study, we explored genetic variation between the extreme populations for both single nucleotide polymorphisms

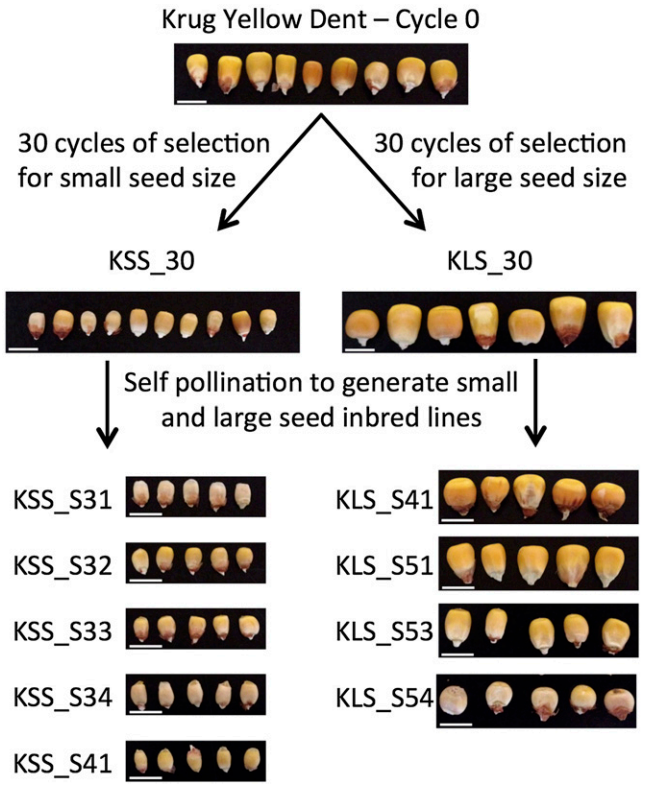


Figure 1 Phenotypic response to selection for large and small seed size. Thirty cycles of divergent selection for seed size was conducted from the base population Krug Yellow Dent to generate KLS_30 (selected for larger seeds) and KSS_30 (selected for smaller seeds). Inbred lines were generated from both KLS_30 and KSS_30 by self-pollinating random plants from each population for at least five generations.

(SNPs) and copy-number variation (CNV), identified regions under selection during the long-term selection experiment, and compared these results to naturally occurring genetic variation in maize for seed weight to elucidate the genetic architecture of seed size in an important cereal crop.

Materials and Methods

Plant material, nucleic acid isolation, and SNP genotyping

The open pollinated maize population Krug Yellow Dent (PI 233006) and its derivatives were evaluated in this study. Thirty cycles of divergent mass selection for seed size were conducted to generate KLS_30 (selected for large seed size; PI 636488) and KSS_30 (selected for small seed size; PI 636489) (Odhiambo and Compton 1987; Russell 2006). Briefly, in each cycle of selection, 1200 to 1500 plants from each divergently selected population were grown in separate isolation blocks, ears with the consistently largest or smallest seeds were selected (minimum of 100 ears per population), and an equal number of seeds from each ear was bulked to constitute the population for the next cycle of selection. Additionally, inbred lines were generated from both KLS_30 and KSS_30 by self-pollinating random plants

from each population for at least five generations without selection for seed characteristics (Figure 1; KLS_S41, KLS_S51, KLS_S53, KLS_S54, KSS_S31, KSS_S32, KSS_S33, KSS_S34, and KSS_S41).

Plants from the three populations and the nine inbred lines were grown under greenhouse conditions (27°/24° day/night and 16 /8 hr light/dark). Leaf tissue was harvested from 48 individuals from each population and the nine inbred lines. DNA was extracted using the cetyl (trimethyl)ammonium bromide (CTAB) method (Saghai-Maroof *et al.* 1984). Genotyping was performed by Pioneer Hi-Bred International (Johnston, IA) on individual DNA samples using an Illumina BeadArray 768 SNP assay (Jones *et al.* 2009).

Library construction and sequencing

Three equimolar pools of total DNA were created from the 48 individuals within each population (Krug Yellow Dent, KLS_30, and KSS_30). Libraries were prepared using the Illumina protocol (San Diego, CA) with a target insert size of 270 bp. Sequencing was performed at the Joint Genome Institute (Walnut Creek, CA) using an Illumina HiSeq (San Diego, CA) to generate 2 × 100 nucleotide paired-end sequence reads. Sequence reads are available in the National Center for Biotechnology Information Sequence Read Archive study accession no. SRP013705. The FastQC program (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) was used to examine sequence quality. Reads with insufficient quality were removed from downstream analyses.

Genomic sequence analysis

Genomic reads were cleaned using the FASTX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html) and mapped using Bowtie v. 0.12.7 (Langmead *et al.* 2009) according to previously described methods (Beissinger *et al.* 2014) with the exception that reads were mapped only as single-end reads using the "SE pipeline." For each population, valid alignments were processed using SAMtools v. 0.1.12a (Li *et al.* 2009) as previously described (Beissinger *et al.* 2014) to identify polymorphic positions and determine frequencies of each nucleotide at each position.

It is possible that some of the polymorphic loci were actually the result of multiple copies of a genomic region in one or more of the individuals mapping to a single locus in the B73 reference sequence. As such, a high confidence set of SNPs was identified by placing a constraint on coverage at each position, requiring coverage ±2 standard deviations of the mean across the populations and a minimum coverage of 20× to ensure accurate estimation of allele frequencies in the populations (20× and 79× coverage). After this filtering, 3,090,214 high-confidence SNPs were retained.

A permutation test was used to determine the probability of the difference in observed mean minor allele frequency (MAF) between the SNPs that were fixed in both populations in the same direction and the SNPs that were fixed in both populations in opposite directions. The set of 447,328 SNPs that were polymorphic in Krug Yellow Dent and reached fixation in both populations (in the same and opposite direction) were randomly shuffled 10,000 times and the number of instances when the difference in mean MAF exceeded the empirical observation was recorded.

The distribution of read-depth variation across the genome was used as a proxy to evaluate CNV between the three populations. Read depth was determined for 5-kb windows. Copy-number variation windows were defined as having an absolute value greater than two for the number of standard deviations away from the mean in KLS_30 minus the number of standard deviations away from the mean in KSS_30. Graphical images were generated using R v. 2.13.2 (R Development Core Team 2014) and Circos v. 0.56 (Krzywinski *et al.* 2009).

Comparative genomic hybridization

Comparative genome hybridization (CGH) was performed on the nine inbred lines generated from the KLS_30 and KSS_30 populations and the B73 maize reference inbred line using a previously described microarray design (Eichten *et al.* 2013; GEO Platform GPL15621) and hybridization methodology (Swanson-Wagner *et al.* 2010). Pair files exported from NimbleScan (Nimblegen Inc.) were normalized to correct for signal variations within and between arrays using variance stabilization and calibration (vsn; Huber *et al.* 2002). Normalized samples were exported as $\log^2$(sample/B73 reference) values. The nine individual samples, as well as contrasts between the average KLS and KSS inbred values, were processed into segments via DNAcopy (Venkatraman and Olshen 2007) to identify regions exhibiting CNV. Segments were filtered to require a 0.7-fold change between the two samples to be classified as a CNV.

Estimating effective population size

Three methods were used to measure the effective population size throughout selection in the two directional selection experiments. The first method was based on population demographics as previously described (Crow and Kimura 1970), based on the relationship $N_e = (4N_mN_f)/(N_m + N_f)$, where N_m and N_f are the number of mating males and females, respectively. Next, an estimate was made on the basis of a temporal assessment of molecular markers. Effective population size based on the Illumina BeadArray SNPs was estimated using the equation $N_e = 1/2(1 - \sqrt[t]{H_t/H_0})$, where H_t and H_0 are the mean levels of heterozygosity in the tth and 0th generation, respectively (Crow and Kimura 1970). A third analysis was conducted on the basis of linkage disequilibrium (LD) among the same set of SNPs. Unlike the previous two approaches, this technique allows the estimation of N_e for each of the three populations independently and also provides a confidence interval around the estimates. The program LDN*e* (Waples and Do 2008) was used for this analysis. All SNPs with allele frequencies ≥0.05 were included, and confidence intervals were estimated using the JackKnife approach.

Simulations of drift

Two sets of drift simulations that assumed linkage equilibrium were conducted using R v. 2.15.3 (R Development Core Team 2014). The first set was based on population demography, mimicking the selection protocol exactly. The second set assumed equal males and females and assumed the N_e values estimated from LDN*e* (Waples and Do 2008), which suggested an effective population size of ~14 males and 14 females for both KLS_30 and KSS_30. In both cases, 1000 simulations were conducted. For each simulation, 1,000,000 polymorphic SNPs were sampled, with replacement, from observed polymorphic cycle zero SNPs to create a simulated base population with 1,000,000 allele frequencies. Then, binomial sampling was conducted to mimic 30 generations of drift with the prescribed population size, to generate simulated KLS_30 and KSS_30 populations. Binomial sampling of 96 alleles from each of the three simulated populations (Krug Yellow Dent, KLS_30, and KSS_30) was conducted to mimic sampling individuals to be sequenced. Sequencing was simulated by binomial sampling, for each SNP, the number of reads that were actually sequenced for that SNP in the experiment. SNPs that were simulated to be fixed in the same direction in all three populations were removed, since our SNP calling protocol would not have identified these as polymorphic. The mean percentage of SNPs fixed in opposing directions between KLS_30 and KSS_30 was calculated for each set of simulations, as well as 95% intervals.

Scan for selection

A genome-wide scan for selection was conducted. The use of pooled sequencing prevented estimation of LD in the populations, making accurate simulations to establish precise significance levels impossible. Instead, a window-based scan was used to classify genomic regions as empirically divergent or not divergent. The most divergent sites represent candidates for selection. This approach has been implemented in other studies that have documented strong selection and dramatic phenotypic changes (Beissinger *et al.* 2014) as is the case in this study.

The high confidence set of SNPs described above was further filtered to include only biallelic SNPs (2,944,220 SNPs included). Minor allele frequency as defined in Krug Yellow Dent was calculated in all three populations using a maximum-likelihood estimate. A sliding window approach was used to evaluate divergence between the populations, as there is a substantial sampling error inherent to pooled sequencing.

For each SNP, three F_{ST} values were calculated, corresponding to comparisons between Krug Yellow Dent and KLS_30, Krug Yellow Dent and KSS_30, and KLS_30 and KSS_30. F_{ST} was calculated using a method assuming a large sample size, given by

$$\widehat{F_{ST}} = \frac{s^2}{\bar{p}(1-\bar{p}) + s^2/r},$$

where $\bar{p}$ is the mean allele frequency across populations, s^2 is the variance of allele frequency between populations, and r is the number of populations (Weir and Cockerham 1984). F_{ST} values were averaged over 25-SNP sliding windows, centered on each SNP in turn, to reduce sampling error. This approach assumes that SNP density is high enough that regions under selection will contain multiple SNPs and thus exhibit large F_{ST} values after averaging.

Outlying SNPs, for which the window-averaged F_{ST} value exceeded a 99.9% or 99.99% empirically determined threshold, were identified. These outlier threshold levels were not chosen to represent a specific level of significance; rather they provide candidates for strong (99.9%) or extremely strong (99.99%) selection. To define regions that were putatively under selection, single or adjacent SNPs that displayed an outlying window-averaged F_{ST} value were first identified. Then, if any other SNPs within 5 Mb displayed an outlying window-averaged F_{ST} value, the selected region was extended to include these SNPs. This process was repeated until no significant SNPs were found within 5 Mb of the up- or downstream region boundaries. To ensure that region boundary declarations were conservative, we extended the boundaries to include all of the SNPs in the windows for those SNPs within the extended selection regions (Supporting Information, Table S1 and Table S2).

A map of centimorgans per megabase in the intermated B73 × Mo17 (IBM) population (Lee *et al.* 2002) was previously estimated (Liu *et al.* 2009). This map was used to approximate the relative levels of recombination across the genome of the Krug long-term selection populations. This analysis assumes that recombination hot and cold spots are likely similar across populations. Each of the F_{ST} -based regions that exceeded the 99.9% outlier level was assigned a value for centimorgans per megabase according to the IBM map. The Pearson correlation between region size and region centimorgans per megabase was tested. This was conducted for every region identified, as well as for each comparison separately (KLS_30 *vs.* KSS_30, Krug Yellow Dent *vs.* KLS_30, Krug Yellow Dent *vs.* KSS_30).

Evaluation of natural variation

The maize nested association mapping (NAM) population (Yu *et al.* 2008; Mcmullen *et al.* 2009) was used to evaluate natural variation for seed weight, excluding the two sweet corn families (IL14H and P39). In total, 4196 recombinant inbred lines (RILs) from the non-sweet corn families were used in this study.

The NAM RILs were grown at four locations in 2006 (Clayton, NC; Aurora, NY; Homestead, FL; and Ponce, PR) and at one location in 2007 (Clayton, NC). At each location, a single replicate with checks was planted in an augmented design as previously described (Buckler *et al.* 2009). Seed weight was measured as the weight of 20 representative seeds from two self-pollinated plants per plot. The best linear unbiased predictions (BLUPs) of RILs across environments were calculated with ASREML v. 2.0 software (Gilmour *et al.* 2006) as previously described (Hung *et al.* 2012). The BLUPs were used for subsequent analysis.

Joint linkage mapping was performed according to previously described methods (Buckler *et al.* 2009) using 1106 SNP markers (McMullen *et al.* 2009). Based on 1000 permutations, the appropriate *P*-value for inclusion of a marker in the joint linkage mapping was determined to be 2.03×10^{-6}. Genome-wide association studies (GWAS) were performed using 1.6 million SNPs from the maize HapMap v. 1 project (Gore *et al.* 2009) projected onto the NAM RILs as previously described (Tian *et al.* 2011). Briefly, SNP associations were tested for each chromosome separately. RIL residual values from a model containing QTL identified by the joint linkage model outside of the test chromosome were used as the input phenotype values to GWAS for a particular chromosome. Forward regression was performed on one chromosome at a time, and significance thresholds for each chromosome were determined by 1000 permutations (range from 6.6×10^{-9} to 7.3×10^{-8}). Additionally, the resampling model inclusion probability (RMIP) method for GWAS was performed as previously described (Tian *et al.* 2011). For this method, 80% of the RILs from each family were randomly selected without replacement and forward regression was performed. This method was repeated 100 times, and SNPs that were selected in the regression model in five or more subsamples were considered significant (RMIP $\geq$ 0.05).

Results

Effective population size in the Krug Yellow Dent long-term artificial selection experiment

In the original selection experiment, ~1200 plants per cycle were evaluated, from which ~100 females were selected (Odhiambo and Compton 1987; Russell 2006). Assuming random mating throughout the experiment, the effective population size based on population demographics was estimated to be ~369 for both KLS_30 and KSS_30. Using the 768 SNP markers on individual plants, the effective population size based on observed reductions in heterozygosity was estimated to be 76 and 312 for KSS_30 and KLS_30, respectively. Estimates based on LD for each population using the 768 SNP markers were 33.5 (95% confidence interval, 32.8–34.3) for Krug Yellow Dent, 29.0 (28.3–29.7) for KSS_30, and 27.6 (27.0–28.2) for KLS_30. The differences in N_e resulting from the heterozygosity-based method compared to the LD method may result because the heterozygosity method does not incorporate information about N_e in the base population (Krug Yellow Dent), while the LD method depicts it as relatively low. Still, only a slight reduction in N_e was observed between the base and selected populations based on the LD method, which is in general agreement with the fact that larger N_e was estimated according to reductions in heterozygosity.

Single nucleotide polymorphism detection and estimates of allele frequencies

We generated a total of 462 Gb of sequence across the three population pools, with theoretical coverage of 71.1×, 48.3×, and 81.6× for Krug Yellow Dent, KLS_30, and KSS_30, respectively. The maize genome is highly repetitive (Schnable *et al.* 2009) and as such it is not possible to map to the majority of the genome when a sequence read is required to have a unique alignment. Despite this characteristic, coverage of 58–63% of the base pairs in the reference sequence across the three populations was observed, and 7–18% of the genome had >20× coverage (Table S3).

The result of 30 generations of divergent selection is reflected in probability density curves of the major allele frequency, where the density at a major allele frequency of one is greater in KLS_30 and KSS_30 relative to Krug Yellow Dent (Figure 2A). Interestingly, for 25% of the polymorphic loci, alleles were observed in KLS_30 or KSS_30 that were not present in Krug Yellow Dent (Figure S1). Most likely this is the result of alleles that were present at too low a frequency in Krug Yellow Dent to be detected through sampling of 96 gametes and subsequent sequencing of only a subset of these. Alternatively, this could be the result of accidental introgression or mutations that arose during the experiment and were selected upon.

Identification of regions that exhibit substantial divergence

The genome was scanned to identify candidate regions under selection using an outlier-based approach. Regions exceeding either the 99.9 or 99.99% levels of the empirical distribution were identified. Comparisons were made between Krug Yellow Dent and KLS_30, Krug Yellow Dent and KSS_30, and KLS_30 and KSS_30 (Figure 3, Figure S2, Table S1, and Table S2). A window-based approach was implemented to minimize the effect of sampling error incurred through pooled sequencing while retaining signal from selected regions due to the relatively dense SNP markers that were identified. However, in regions with small selection signatures or relatively low SNP density, this approach can result in undetected selection signatures.

In total, 94 regions that encompass 147.2 Mb (6.4%) of the maize v. 2 reference genome sequence (including *N*'s) were identified as divergent at the 99.9% outlier level and these included 23 regions (25.1 Mb) at the 99.99% level (Table S1 and Table S2). The selected regions contained 2423 and 305 annotated genes at the 99.9% and 99.99% levels, respectively. Among the regions identified at the 99.9% level, 63 were identified in KLS_30 and 27 in KSS_30, based on comparison with Krug Yellow Dent, while direct comparison of KLS_30 and KSS_30 identified 23 regions. Considerable overlap of regions identified in the three comparisons was observed (Figure 4).

Based on a previously described recombination map (Liu *et al.* 2009), no significant correlation between the size of selected regions and the expected relative level of recombination in the corresponding area of the genome was observed (Figure S3). This was the case for regions identified from Krug Yellow Dent *vs.* KLS_30 (*P*-value = 0.2152), Krug Yellow Dent *vs.* KSS_30 (*P*-value = 0.4081), KLS_30 *vs.* KSS_30 (*P*-value = 0.9142), and all identified regions at once (*P*-value = 0.2276).

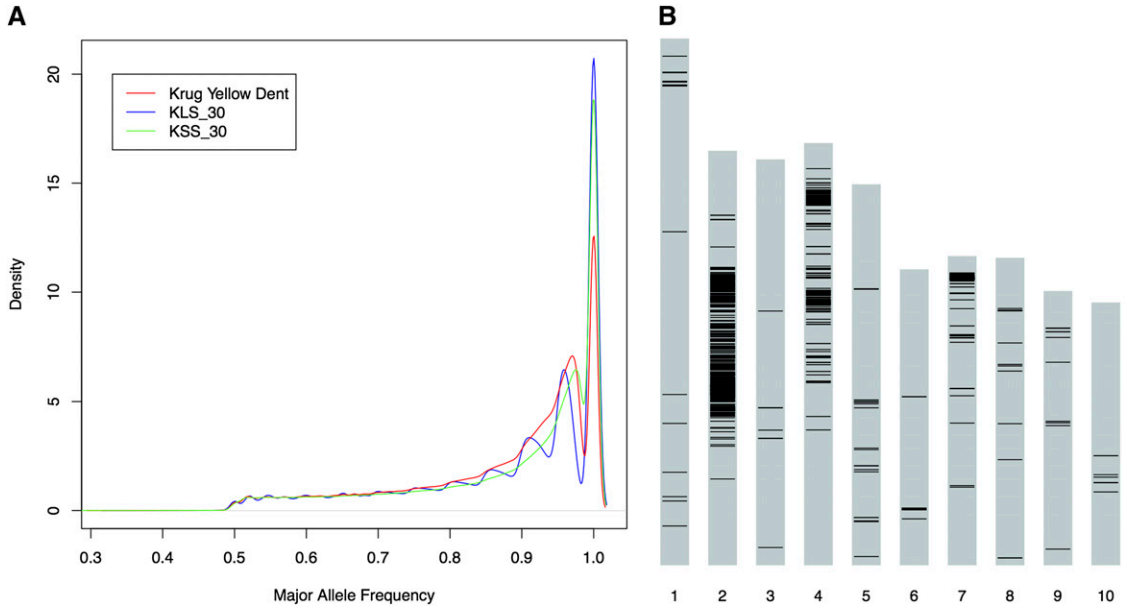


Figure 2 SNP diversity in Krug Yellow Dent, KLS_30, and KSS_30. (A) Probability density function of major allele frequencies for each population based on 3,090,214 high-confidence SNPs with at least 20× coverage and no more than 79× coverage. The area under each curve equals one. (B) Distribution of SNPs that reached fixation in both KLS_30 and KSS_30 with opposing alleles in the extreme populations, reflecting the divergent selection.

However, even though no significant correlation was observed, the largest region located on chromosome 2, which displayed evidence of selection based on all three comparisons, did fall in an area of very limited recombination.

Across the three comparisons, the number of genes within 5 kb of selected regions ranged from 0 to 233 with a mean of ~27 (Table S1 and Table S2). However, a small number of large candidate regions skewed this value upward. Interestingly, candidate regions for selection were observed on chromosome 2 and 4 in the KSS_30 population (Figure S2), and the heterozygosity-based estimate of effective population size was lower in KSS_30 compared with KLS_30. It is unknown, however, if an undocumented bottleneck resulted in these large candidate regions of selection, or if large sweeps caused a bottleneck to occur in the population.

In contrast to the mean number of genes per region, the median number of genes within the identified regions was six, and 28 regions contained only one or zero genes within the region. Candidate genes were identified within some of the regions. For example, region 20 on chromosome 7 (Figure 3 and Table S2) contained *o2*, which is known to regulate expression of genes encoding 22-kDa zein proteins (Schmidt *et al.* 1990, 1992) and is expressed almost exclusively in developing seed tissue with the highest expression levels observed in endosperm tissue (Sekhon *et al.* 2011). While SNPs from this study within *o2* did not show evidence of changes in allele frequency, significant differences in expression were observed throughout development between KLS_30 derived inbred lines and KSS_30 derived inbred lines (Figure S4) (Sekhon *et al.* 2014).

In a previous study, gene coexpression network modules that distinguish KLS_30 and KSS_30 derived inbred lines were identified, one of which was enriched with cell-cycle genes (Sekhon *et al.* 2014). Nineteen genes within 14 different genomic regions identified at the 99.9% level were within this cell-cycle-enriched module (Table S4). One of these genes (GRMZM2G069078) has previously been shown to have an effect on seed development in the maize UniformMu mutant population (McCarty *et al.* 2005; Hunter *et al.* 2014). Interestingly, expression patterns in the KLS_30 and KSS_30-derived inbred lines indicate differences in developmental timing, with the gene expressed longer in the KLS_30 inbred lines (Figure S5)(Sekhon *et al.* 2014).

Four genes within our identified regions were within another gene coexpression network module that was enriched in zein proteins from the same network analysis (Sekhon *et al.* 2014). One of these genes was annotated as a starch binding domain containing protein (GRMZM2G161534; genomic region 70, chromosome 6; Table S1) and one as a 22-kDa alpha zein protein 21 (GRMZM2G397687; selective sweep 36, chromosome 4; Table S1).

A large number of single nucleotide polymorphisms reached fixation in the selected populations

In total, 1,111,384 loci that were polymorphic in Krug Yellow Dent reached fixation in KLS_30 and/or KSS_30 (Figure S1). Many of these observed positions could be due to sampling of alleles that were in low frequency in the base population and were sampled in only one of the selected populations. There was, however, a subset of these SNPs (2729; 0.088% of analyzed SNPs) that reached fixation in both KLS_30 and KSS_30 with opposing fixed alleles between the two extreme populations that were distributed across the 10 chromosomes (Figure 2B). A large number

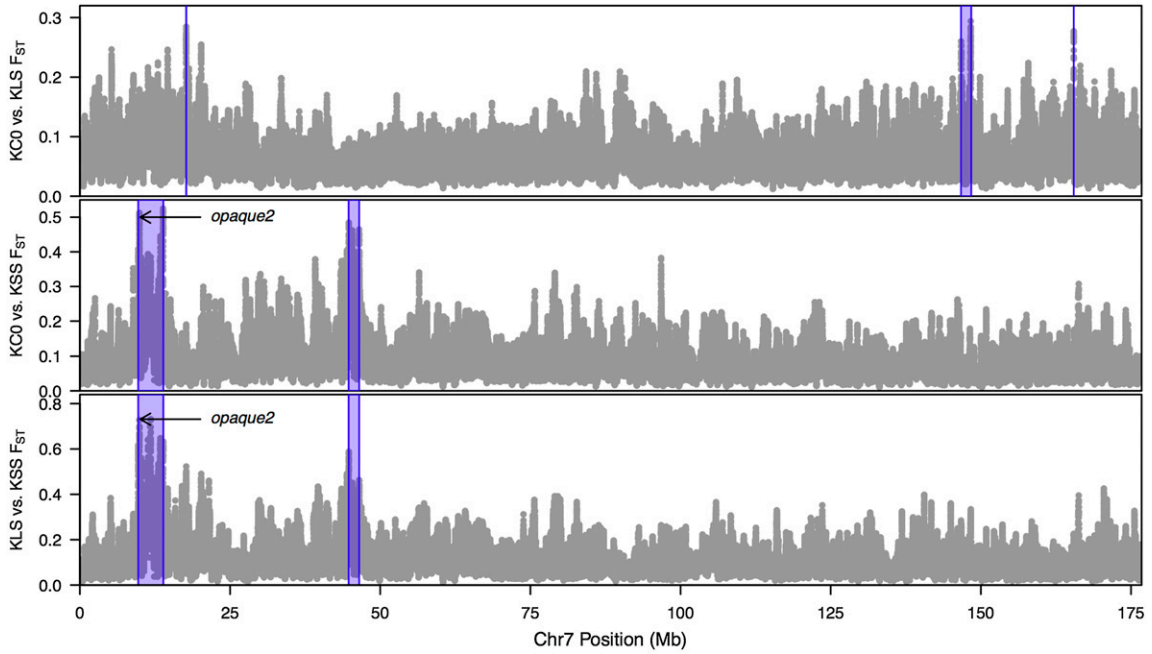


Figure 3 Window-averaged F_{ST} values for the SNPs on chromosome 7. F_{ST} values were calculated using a 25-SNP sliding window approach for the biallelic SNPs. Comparisons were made between Krug Yellow Dent and KLS_30, Krug Yellow Dent and KSS_30, and KLS_30 and KSS_30. Purple areas indicate candidate regions under selection at the 99.9% level. Plots for all chromosomes with 99.9 and 99.99% threshold values are available in Figure S1. KC0, Krug Yellow Dent; KLS, KLS_30; KSS, KSS_30.

of the oppositely fixed SNPs were clustered near the centromere on chromosome 2 and on the short arm of chromosome 4 (Figure 2B). As was expected, significant overlap was observed with the candidate regions identified by the outlier-based scan of the genome described above (Figure 4). Interestingly, however, small regions of fixation, in some cases a single oppositely fixed SNP, that did not overlap with the regions identified using the window-based outlier-based approach were observed. However, in many cases the oppositely fixed SNPs were consistent with allele frequency changes at surrounding loci that simply had not yet reached fixation.

The MAF of SNPs that were fixed in opposite directions was substantially higher (mean MAF 0.233) than that observed for SNPs that reached fixation in only one population (mean MAF 0.175) and for all SNPs in the base population (mean MAF 0.175; Figure S6). Permutation analysis showed a significant difference in the mean MAF between the two classes of fixed SNPs (fixed in both populations in the same or opposite directions; P-value = 0.0001). The probability of differential fixation can be calculated as $P(1 - P)$, where P is the probability of fixation. Based on this equation, differential fixation becomes more likely as MAF approaches 0.5. Thus, the observed SNPs that were fixed in opposite directions likely resulted, at least in part, from drift during the 30 cycles of selection.

Simulations were also conducted to determine the expected number of SNPs to be fixed in opposite directions due to drift alone. The mean percentage of opposite-fixed SNPs based on simulations with effective population size determined according to demography was 2.8×10^{-6}% (95% interval: 0.0%–1.05×10^{-4}%), which is substantially fewer than the observed percentage. It should be noted, however, that the mean percentage of opposite-fixed SNPs based on simulations with effective population size determined by LDN*e* (Waples and Do 2008), which provided the lowest estimate of N_e among the methods utilized, was 0.7% (95% interval 0.77–0.81%).

Copy-number variation was highly prevalent between KLS_30 and KSS_30

Using read-depth variation as a proxy for CNV, 57 variable 5-kb windows were identified between the selected populations (Figure 5A and Table S5). Some of the CNV regions contained multiple significant windows in close proximity (Figure 5B), while others had only a single window above the background noise (Figure 5C). Interestingly, CNV regions that did not contain any annotated gene models and may be involved in regulation of gene expression were identified.

The putative CNV regions from read-depth variation were identified from a pool of 48 individuals. Thus, these may represent regions that had modest changes in copy number in many individuals or extreme changes in copy-number variation in a small number of individuals. To provide perspective on the basis of the CNV regions identified from the pooled resequencing experiment, CGH was performed on individual inbred lines derived from the populations. From the CGH, 479 regions were identified with variation between the average of the large and small seeded inbred lines derived from the extreme populations (Figure 1 and Table S6). Notably, four of the read-depth variants were also

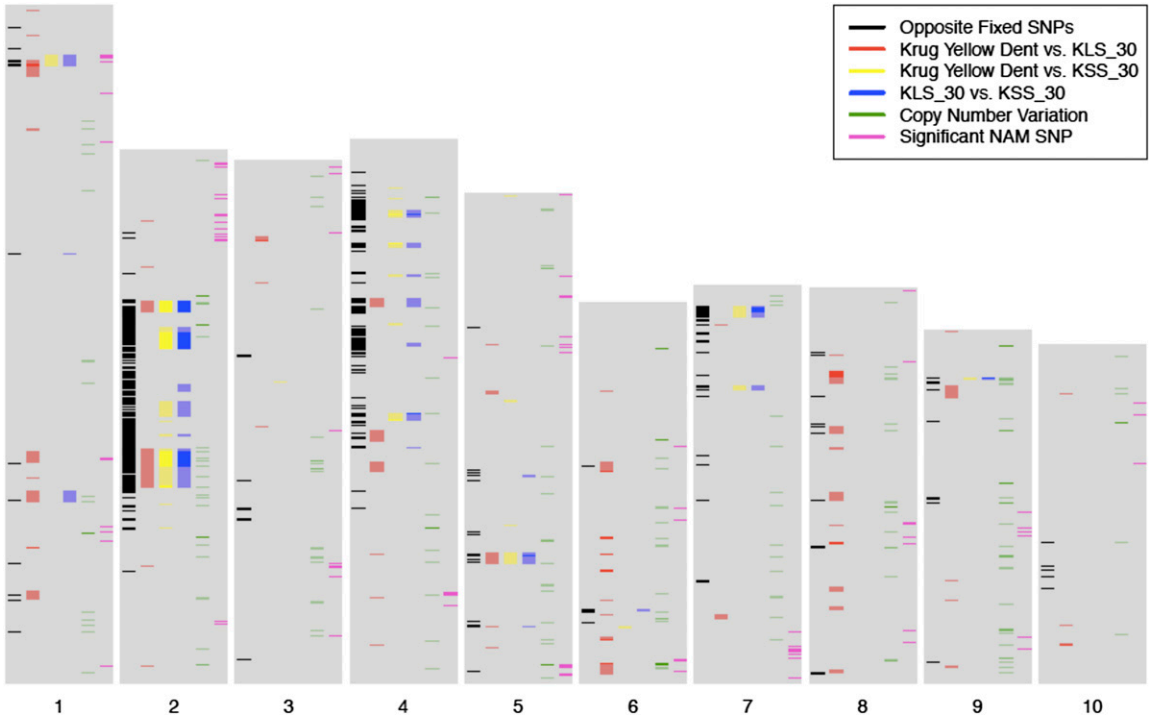


Figure 4 Distribution of genetic variation in the Krug Yellow Dent divergent long-term selection experiment for seed size and quantitative trait loci for seed weight in the maize nested association (NAM) population along the 10 maize chromosomes. Opposite fixed SNPs are those that have reached fixation in both KLS_30 and KSS_30 with opposing alleles. Krug Yellow Dent *vs.* KLS_30, Krug Yellow Dent *vs.* KSS_30, and KLS_30 *vs.* KSS_30 show candidate genomic regions under selection observed in the various comparisons at the 99.99% level (opaque colors) and 99.9% level (transparent colors). Opaque green bars indicate copy-number variation (CNV) regions that were identified from pooled resequencing data from the populations and transparent green bars indicate regions that were identified from comparative genome hybridization (CGH) with inbred lines derived from KLS_30 and KSS_30. Significant NAM SNPs include SNPs identified using both joint linkage analysis and genome wide association studies.

identified using the CGH method (Figure 5A), which significantly exceeds the overlap expected by chance (Figure S7). Using the two methods, a total of 532 CNV regions were identified between the extreme populations (53 unique to the read depth variants, 475 unique to the CGH CNVs, and 4 overlapping regions).

Of the 532 CNV regions identified, 148 contained or overlapped at least one gene annotated in the maize v. 2 reference sequence. Of the CNV regions containing annotated genes, 15 contained genes important for photosynthetic activity including photosystem I and photosystem II proteins and a RuBisCO large-chain protein. Interestingly, previous phenotypic evaluation of these populations revealed variation for mature plant dry weight in addition to seed size (Sekhon *et al.* 2014). Eight cell-cycle genes, such as cyclin protein-coding genes, were also present in the CNV regions. As discussed above, previous comparison of whole transcriptomes between the KLS_30 and KSS_30-derived inbred lines identified a gene coexpression module that differentiated the inbred lines and contained a large number of cell-cycle-related genes (Sekhon *et al.* 2014). Notably, three of the genes identified in regions with CNV were contained in this module including one annotated as an auxin-independent growth promoter on chromosome 5.

Overlap was also observed between the CNV regions and the regions that were identified as the most likely to be affected by selection based on SNP allele frequencies. However, the overlap exceeded only that expected by chance for the CNV regions identified by CGH (Figure S8). Across the 94 regions that were identified at the 99.9% level, 29 were within 5 kb of a CNV region identified by CGH (28) or sequence depth (2). Of particular interest, region 71 on chromosome 6 overlapped with both CGH and sequence-depth-identified CNV regions, and this region also contained three genes that were in the cell-cycle-enriched gene co-expression module described above (Table S4) (Sekhon *et al.* 2014). Additionally, two of the three CNV regions on chromosome 2 were within the SNP divergently fixed regions (Figure 2B).

Natural genetic variation for seed weight validates regions identified in the Krug Yellow Dent selection experiment

To compare artificial selection in the Krug long-term selection experiment with natural variation for seed size, 20-kernel seed weight, a trait highly correlated with seed size (Peng *et al.* 2011), was evaluated in the maize NAM population (Yu *et al.* 2008; McMullen *et al.* 2009). Briefly, the NAM population includes 25 RIL families, each with B73 as a common reference parent. The 25 NAM founders were selected to maximize diversity from a worldwide collection of maize

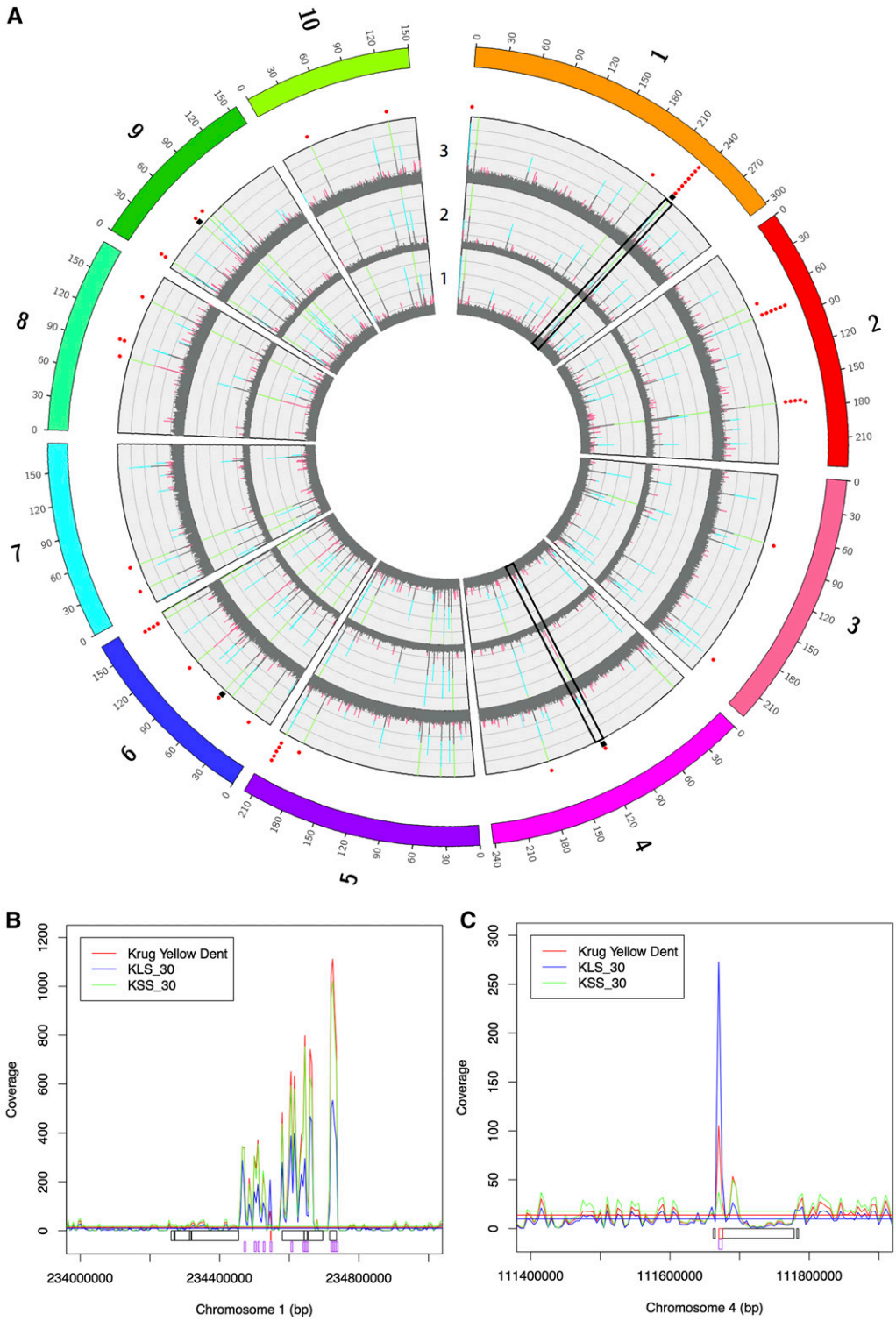


Figure 5 CNV in Krug Yellow Dent, KLS_30, and KSS_30 based on read-depth variation and comparative genome hybridization (CGH). (A) Distribution of average read depth in 5-kb windows for Krug Yellow Dent (track 1), KLS_30 (track 2), and KSS_30 (track 3). Pink indicates a window that is >1 SD above the mean for the given population, aqua indicates a window that is >2 SD above the mean for a given population, and green indicates a window that has >250× read depth and extends beyond the chart. Red dots outside of track 3 show windows with evidence of CNV based on read depth (defined as the number of SD away from the mean in KLS_30 minus the number of SD away from the mean in KSS_30 being greater than two). Black squares outside of track 3 show CGH probes with significant CNV between KLS_30 and KSS_30-derived inbred lines that are concordant with sequence-based CNV regions at the population level. (B) Close-up of a significant CNV region on chromosome 1. (C) Close-up of a significant CNV region on chromosome 4. In both B and C, black boxes indicate CGH regions that do not show CNV, red boxes indicate CGH regions that show CNV, and purple boxes indicate 5-kb read-depth variation windows.

inbred lines based on microsatellite markers (Liu *et al.* 2003; Flint-Garcia *et al.* 2005; Yu *et al.* 2008) and are thus a good representation of natural variation in maize inbreds. The two sweet corn families in the NAM population were excluded from the analysis due to their extreme seed weight phenotypes. The parents of the included families were both genotypically and phenotypically diverse, with 20-kernel seed weights ranging between 2.18 and 5.32 g. In comparison, the average 20-kernel seed weight for the KSS_30 and KLS_30 populations was previously reported to be 1.96 and 9.35 g, respectively (Sekhon *et al.* 2014).

Using joint linkage analysis, 18 QTL peaks were identified for seed weight (Table S7), which accounted for 60% of the total phenotypic variation, with the range in additive

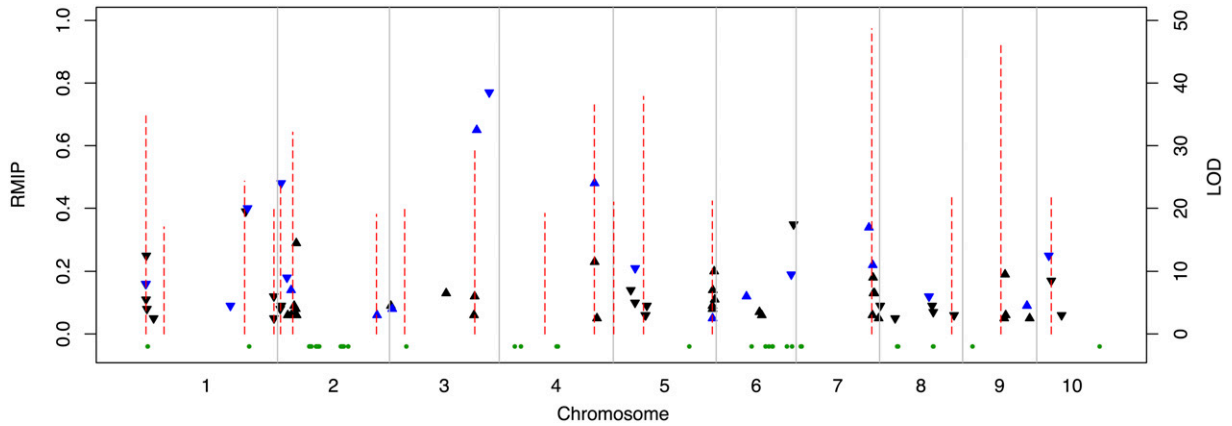


Figure 6 Position and magnitude of genetic variation underlying natural variation for seed weight in the maize NAM population. Red dotted lines depict significant QTL peaks based on joint linkage analysis (scale log of odds, LOD). Triangles depict associations identified from GWAS using the subsampling method (resampling model inclusion probability, RMIP ≥ 0.05). Triangles pointing upward indicate a positive effect and triangles pointing downward indicate a negative effect relative to B73. Blue triangles indicate associations detected using the subsampling and forward regression methods (scale RMIP). Green dots indicate selective sweeps observed in the Krug long-term selection experiment at the 99.99% outlier level.

allelic effect size between −0.012 and 0.013 g per 20 kernels. Overlap was observed between seed weight and seed composition QTL identified in a previous study (starch, 9 QTL; protein, 7 QTL; oil, 7 QTL) that used the same germplasm (Cook *et al.* 2012b), providing additional evidence that seed composition likely contributes to seed size and weight. Single forward regression GWAS using the 1.6 million SNPs from the HapMap v. 1 data set identified 21 SNPs associated with seed weight (Table S8). The RMIP GWAS method using the same HapMap v. 1 data set identified 76 SNPs associated with weight (Table S9), which validated 20 of the 21 SNPs from the single forward regression GWAS model. In total, 74 regions of the genome were associated with seed weight based on joint linkage analysis and GWAS in the NAM population when allowing overlapping regions to be within 500 kb of an adjacent significant SNP (Figure 6, Table S7, Table S8, and Table S9).

Overlap was observed between the variable regions identified in the Krug Yellow Dent divergent selection experiment and the regions identified in NAM, in terms of the read-depth-based CNV regions (6 NAM SNPs), CGH-based CNV regions (25 NAM SNPs), and selective sweeps (12 NAM SNPs) when requiring SNPs to be within 500 kb of a variable region (Figure S8). For both CNV detection methods, this level of overlap exceeded the number expected by chance (Figure S7). Of particular interest was overlap with the large CNV region on chromosome 1 that was detected by both read-depth analysis of the extreme populations and CGH analysis of the population-derived inbred lines (Figure 5B). However, no obvious candidate genes were identified in either the CNV region or in the gene containing the significant NAM SNP. The level of overlap with the regions that exceeded the outlier threshold did not exceed the number of overlapping regions expected by chance with the selective sweeps. This could indicate the presence of many unique regions of the genome underlying the phenotypic variation observed within each population or it could reflect random false positives observed in each population.

Discussion

Cereal crops, including maize, are an important food source worldwide. Understanding the genetic architecture of grain yield and yield component traits is important to producing sufficient food to feed the human population. The populations derived out of the Krug long-term selection experiment (Odhiambo and Compton 1987; Russell 2006) provided a powerful tool for identifying regions of the genome-controlling seed weight and grain yield. The relatively large effective population size that was maintained throughout the experiment, as well as the divergent populations, allowed for separation of selection and drift effects. By resequencing pooled individuals from the base and selected populations, we were able to identify regions of the genome that were altered in response to selection for seed size.

Our observation of no significant relationship between recombination rate and the size of F_{ST} -based regions has interesting implications from an evolutionary standpoint. Generally speaking, selection sweeps can be classified as "hard sweeps," for which a mutation arises and is immediately beneficial in the population (Maynard Smith and Haigh 1974), and "soft sweeps," for which standing variation becomes beneficial due to a change in selection pressure (Hermisson and Pennings 2005). It is unlikely that any type of selection pressure occurred before the artificial selection program began, and because of the limited number of generations of selection, novel mutations affecting the trait are improbable. In an independent maize population subjected to a comparable selection protocol, soft sweeps were predominantly observed (Beissinger *et al.* 2014), and our *a priori* expectation was that mostly soft sweeps had occurred in this study. Unlike the findings by Beissinger

et al. (2014), where most sweeps were classified as soft according to size, a large and relatively continuous distribution of region size was observed in the Krug long-term selection experiment (Figure S3). Additionally, region size in the Krug population did not appear to be controlled primarily by recombination rate. While inconclusive, these results indicate that the populations may have undergone classical hard sweeps, soft sweeps, and a combination thereof.

Some of the regions identified in our current study were small and allowed for candidate genes under selection to be identified. For example, *o2* was contained in one of the selective sweeps and has been extensively studied for its role in endosperm development, namely in regulating expression of genes encoding 22-kDs zein proteins (Schmidt *et al.* 1990, 1992). Additionally, the significant GWAS signal at the end of the long arm of chromosome 2 is <100 kb from the window to which *stt1* was mapped (Phillips and Evans 2011).

Large candidate regions for selection that likely resulted from genetic hitchhiking (Maynard Smith and Haigh 1974) were also observed in this study. For these regions that contained up to 233 genes, extensive genetic dissection and incorporation of multiple sources of evidence will be required to determine the variant and/or variants underlying them. The gene GRMZM2G069078 on chromosome 8 is a prime example where utilizing multiple sources of evidence including selective sweep analysis, gene coexpression network analysis (Sekhon *et al.* 2014), and mutation analysis (Hunter *et al.* 2014) allowed for the identification of a gene that was likely selected in the Krug long-term selection experiment.

Interestingly, there were also regions that contained no annotated genes. It is well documented that variants in noncoding regions can have a large effect on phenotypic variation. For example, variants in the maize *Vgt1* region, which is 70 kb upstream of the *ZmRap2.7* gene, were shown to be associated with a flowering time quantitative trait locus (Salvi *et al.* 2007; Ducrocq *et al.* 2008). It is also possible that genes are present in the reference sequence that were not annotated, are present in the reference inbred line B73 yet absent in the assembly, which has been documented to be incomplete (Schnable *et al.* 2009; Lai *et al.* 2010; Hansey *et al.* 2012; Hirsch *et al.* 2014), or are dispensable genes that are absent from the reference inbred line, but are present at some frequency within the Krug populations.

Previously extensive CNV has been shown across diverse maize inbred lines (Springer *et al.* 2009; Lai *et al.* 2010; Swanson-Wagner *et al.* 2010; Chia *et al.* 2012). It has long been hypothesized that this variation is in part underlying the large phenotypic variation in maize. A recent example of aluminum tolerance was associated with three tandem copies of the MATE1 gene in tolerant lines relative to the sensitive lines that carry only one copy of the gene (Maron *et al.* 2013). Likewise, resistance to the soybean cyst nematode was associated with increased copy numbers of three distinct genes (Cook *et al.* 2012a). In the current study, a large number of regions were identified that have altered copy number between the selected populations, KLS_30 and KSS_30 as estimated by read-depth variation and CGH.

A large number of the genes in the CNV regions were related to photosynthetic activity. Phenotypic evaluation of the KLS_30 and KSS_30 populations revealed variation for mature plant dry weight (Sekhon *et al.* 2014), consistent with the presence of photosynthesis-related genes in the CNV regions. Additionally, a number of cell-cycle-related genes were within the CNV regions. Cell-cycle programs are involved in multiple stages of endosperm development including acytokinetic mitosis, cellularization, cell proliferation, and in the cereals, endoreduplication (Kowles *et al.* 1990; Sabelli and Larkins 2009). The presence of cell-cycle genes within CNV regions in this study provides additional support for a growing body of evidence demonstrating the role of master cell-cycle regulators in endosperm formation, development, and seed and plant size (Sabelli and Larkins 2009; Sekhon *et al.* 2014).

Interestingly, obvious candidate genes were not identified in the CNV region on chromosome 1 that was identified by both read depth and CGH or in the gene containing the significant NAM SNP in close proximity to the region. However, there is a B-type response regulator (GRMZM2G379656) that lies between these two regions. In *Arabidopsis thaliana*, B response regulators have been shown to play a role in plant development including mean rosette diameter and mean seed length through regulation of the cytokinin signaling pathway (Argyros *et al.* 2008). A microarray-based gene expression atlas of 60 tissues from the maize reference inbred line B73 showed expression of this gene in leaf tissue at the V5, V9, V10, and R2 developmental stages across three biological replicates (Abendroth *et al.* 2011; Sekhon *et al.* 2011). Additionally, two of the three endosperm replicates at 20 days after pollination showed expression above background, indicating that this gene may also be important in both vegetative and seed development in maize.

This study provides valuable candidate genes that will be useful in characterizing control of seed weight and grain yield in cereals. The results are consistent with the importance of both cell-cycle regulation and seed composition in observed phenotypic variation for seed size/weight and ultimately grain yield. This study also provides insight into long-term artificial selection in crop plants, supporting the hypotheses of many genes with small effects underlying seed size and a role for noncoding sequences and copy-number variation in contributing to phenotypic response to selection.

Acknowledgments

We are grateful to Dupont–Pioneer Hi-Bred International, Inc., for providing SNP data. This research was performed using the computer resources and assistance of the UW—Madison Center For High Throughput Computing (CHTC) in

the Department of Computer Sciences. The CHTC is supported by UW—Madison and the Wisconsin Alumni Research Foundation and is an active member of the Open Science Grid, which is supported by the National Science Foundation and the U.S. Department of Energy's Office of Science. This work was funded by the Department of Energy (DOE) Great Lakes Bioenergy Research Center (DOE BER Office of Science DE-FC02-07ER64494). The work conducted by the U.S. DOE Joint Genome Institute was supported by the Office of Science of the U.S. DOE under contract no. DE-AC02-05CH11231. T.B. was supported by the University of Wisconsin Graduate School and by a gift to the University of Wisconsin—Madison Plant Breeding and Plant Genetics program from Monsanto.

Literature Cited

Abendroth, L. J., R. W. Elmore, M. J. Boyer, and S. K. Marlay, 2011 Corn growth and development. PMR 1009. Iowa State University Extension, Ames, Iowa

Akey, J. M., G. Zhang, K. Zhang, L. Jin, and M. D. Shriver, 2002 Interrogating a high-density SNP map for signatures of natural selection. Genome Res. 12: 1805–1814.

Argyros, R. D., D. E. Mathews, Y. H. Chiang, C. M. Palmer, D. M. Thibault *et al.*, 2008 Type B response regulators of Arabidopsis play key roles in cytokinin signaling and plant development. Plant Cell 20: 2102–2116.

Austin, D. F., and M. Lee, 1998 Detection of quantitative trait loci for grain yield and yield components in maize across generations in stress and nonstress environments. Crop Sci. 38: 1296–1308.

Beissinger, T. M., C. N. Hirsch, B. Vaillancourt, S. Deshpande, K. Barry *et al.*, 2014 A genome-wide scan for evidence of selection in a maize population under long-term artificial selection for ear number. Genetics 196: 829–840.

Bhave, M. R., S. Lawrence, C. Barton, and L. C. Hannah, 1990 Identification and molecular characterization of shrunken-2 cDNA clones of maize. Plant Cell 2: 581–588.

Buckler, E. S., J. B. Holland, P. J. Bradbury, C. B. Acharya, P. J. Brown *et al.*, 2009 The genetic architecture of maize flowering time. Science 325: 714–718.

Cheng, W. H., E. W. Taliercio, and P. S. Chourey, 1996 The Miniature1 seed locus of maize encodes a cell wall invertase required for normal development of endosperm and maternal cells in the pedicel. Plant Cell 8: 971–983.

Chia, J. M., C. Song, P. J. Bradbury, D. Costich, N. de Leon *et al.*, 2012 Maize HapMap2 identifies extant variation from a genome in flux. Nat. Genet. 44: 803–807.

Cook, D. E., T. G. Lee, X. Guo, S. Melito, K. Wang *et al.*, 2012a Copy number variation of multiple genes at Rhg1 mediates nematode resistance in soybean. Science 338: 1206–1209.

Cook, J. P., M. D. McMullen, J. B. Holland, F. Tian, P. Bradbury *et al.*, 2012b Genetic architecture of maize kernel composition in the nested association mapping and inbred association panels. Plant Physiol. 158: 824–834.

Crow, J. F., and M. Kimura, 1970 *An Introduction to Population Genetic Theory*. Harper & Row, New York.

Ducrocq, S., D. Madur, J. B. Veyrieras, L. Camus-Kulandaivelu, M. Kloiber-Maitz *et al.*, 2008 Key impact of Vgt1 on flowering time adaptation in maize: evidence from association mapping and ecogeographical information. Genetics 178: 2433–2437.

Eichten, S. R., M. W. Vaughn, P. J. Hermanson, and N. M. Springer, 2013 Variation in DNA methylation patterns is more common among maize inbreds than among tissues. Plant Gen. 6: 1–10.

Flint-Garcia, S. A., A. C. Thuillet, J. Yu, G. Pressoir, S. M. Romero *et al.*, 2005 Maize association population: a high-resolution platform for quantitative trait locus dissection. Plant J. 44: 1054–1064.

Fu, J., Y. Cheng, J. Linghu, X. Yang, L. Kang *et al.*, 2013 RNA sequencing reveals the complex regulatory network in the maize kernel. Nat. Commun. 4: 2832.

Gilmour, A., B. Gogel, B. Cullis, and R. Thompson, 2006 *ASReml User Guide Release 2.0*. VSN Intl., Hemel, Hempstead, UK.

Gore, M. A., J. M. Chia, R. J. Elshire, Q. Sun, E. S. Ersoz *et al.*, 2009 A first-generation haplotype map of maize. Science 326: 1115–1117.

Hansey, C. N., B. Vaillancourt, R. S. Sekhon, N. de Leon, S. M. Kaeppler *et al.*, 2012 Maize (*Zea mays* L.) genome diversity as revealed by RNA-sequencing. PLoS ONE 7: e33071.

Hermisson, J., and P. S. Pennings, 2005 Soft sweeps: molecular population genetics of adaptation from standing genetic variation. Genetics 169: 2335–2352.

Hirsch, C. N., J. M. Foerster, J. M. Johnson, R. S. Sekhon, G. Muttoni *et al.*, 2014 Insights into the maize pan-genome and pan-transcriptome. Plant Cell 26: 121–135.

Huber, W., A. von Heydebreck, H. Sultmann, A. Poustka, and M. Vingron, 2002 Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics 18(Suppl. 1): S96–S104.

Hung, H. Y., C. Browne, K. Guill, N. Coles, M. Eller *et al.*, 2012 The relationship between parental genetic or phenotypic divergence and progeny variation in the maize nested association mapping population. Heredity 108: 490–499.

Hunter, C. T., M. Suzuki, J. Saunders, S. Wu, A. Tasi *et al.*, 2014 Phenotype to genotype using forward-genetic Mu-seq for identification and functional classification of maize mutants. Front. Plant Sci. 4: 545.

Johansson, A. M., M. E. Pettersson, P. B. Siegel, and O. Carlborg, 2010 Genome-wide effects of long-term divergent selection. PLoS Genet. 6: e1001188.

Jones, E., W.-C. Chu, M. Ayele, J. Ho, E. Bruggeman *et al.*, 2009 Development of single nucleotide polymorphism (SNP) markers for use in commercial maize (*Zea mays* L.) germplasm. Mol. Breed. 24: 165–176.

Kesavan, M., J. T. Song, and H. S. Seo, 2013 Seed size: a priority trait in cereal crops. Physiol. Plant. 147: 113–120.

Khaled, A. S., V. Vernoud, G. C. Ingram, P. Perez, X. Sarda *et al.*, 2005 Engrailed-ZmOCL1 fusions cause a transient reduction of kernel size in maize. Plant Mol. Biol. 58: 123–139.

Kiesselbach, T. A., 1999 *The Structure and Reproduction of Corn*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Kowles, R. V., F. Srienc, and R. L. Phillips, 1990 Endoreduplication of nuclear DNA in the developing maize endosperm. Dev. Genet. 11: 125–132.

Krzywinski, M., J. Schein, I. Birol, J. Connors, R. Gascoyne *et al.*, 2009 Circos: an information aesthetic for comparative genomics. Genome Res. 19: 1639–1645.

Lai, J., R. Li, X. Xu, W. Jin, M. Xu *et al.*, 2010 Genome-wide patterns of genetic variation among elite maize inbred lines. Nat. Genet. 42: 1027–1030.

Langmead, B., C. Trapnell, M. Pop, and S. L. Salzberg, 2009 Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10: R25.

Lee, M., N. Sharopova, W. D. Beavis, D. Grant, M. Katt *et al.*, 2002 Expanding the genetic map of maize with the intermated B73 × Mo17 (IBM) population. Plant Mol. Biol. 48: 453–461.

Lewontin, R. C., 1962 Interdeme selection controlling a polymorphism in the house mouse. Am. Nat. 96: 65–78.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009 The sequence alignment/map format and SAMtools. Bioinformatics 25: 2078–2079.

Liu, K., M. Goodman, S. Muse, J. S. Smith, E. Buckler *et al.*, 2003 Genetic structure and diversity among maize inbred lines as inferred from DNA microsatellites. Genetics 165: 2117–2128.

Liu, S., C. T. Yeh, T. Ji, K. Ying, H. Wu *et al.*, 2009 Mu transposon insertion sites and meiotic recombination events co-localize with epigenetic marks for open chromatin across the maize genome. PLoS Genet. 5: e1000733.

Maron, L. G., C. T. Guimaraes, M. Kirst, P. S. Albert, J. A. Birchler *et al.*, 2013 Aluminum tolerance in maize is associated with higher MATE1 gene copy number. Proc. Natl. Acad. Sci. USA 110: 5241–5246.

Maynard Smith, J., and J. Haigh, 1974 The hitch-hiking effect of a favourable gene. Genet. Res. 23: 23–35.

McCarty, D. R., A. M. Settles, M. Suzuki, B. C. Tan, S. Latshaw *et al.*, 2005 Steady-state transposon mutagenesis in inbred maize. Plant J. 44: 52–61.

McMullen, M. D., S. Kresovich, H. S. Villeda, P. Bradbury, H. Li *et al.*, 2009 Genetic properties of the maize nested association mapping population. Science 325: 737–740.

Neuffer, M. G., E. H. Coe, and S. R. Wessler, 1997 *Mutants of Maize*. Cold Spring Harbor Laboratory Press, Plainview, NY.

Odhiambo, M. O., and W. A. Compton, 1987 Twenty cycles of divergent mass selection for seed size in Corn1. Crop Sci. 27: 1113–1116.

Oleksyk, T. K., K. Zhao, F. M. De La Vega, D. A. Gilbert, S. J. O'Brien *et al.*, 2008 Identifying selected regions from heterozygosity and divergence using a light-coverage genomic dataset from two human populations. PLoS One 3: e1712.

Pan, D., S. Zhang, J. Jiang, L. Jiang, Q. Zhang *et al.*, 2013 Genome-wide detection of selective signature in Chinese Holstein. PLoS ONE 8: e60440.

Parts, L., F. A. Cubillos, J. Warringer, K. Jain, F. Salinas *et al.*, 2011 Revealing the genetic structure of a trait by sequencing a population under selection. Genome Res. 21: 1131–1138.

Paulis, J. W., and J. S. Wall, 1977 Comparison of the protein compositions of selected corns and their wild relatives, teosinte and *Tripsacum*. J. Agric. Food Chem. 25: 265–270.

Peng, B., Y. Li, Y. Wang, C. Liu, Z. Liu *et al.*, 2011 QTL analysis for yield components and kernel-related traits in maize across multi-environments. Theor. Appl. Genet. 122: 1305–1320.

Phillips, A. R., and M. M. Evans, 2011 Analysis of stunter1, a maize mutant with reduced gametophyte size and maternal effects on seed development. Genetics 187: 1085–1097.

R Development Core Team, 2014 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.

Russell, W. K., 2006 Registration of KLS_30 and KSS_30 populations of maize. Crop Sci. 46: 1405–1406.

Sabelli, P., and B. Larkins, 2009 The contribution of cell cycle regulation to endosperm development. Sex. Plant Reprod. 22: 207–219.

Sabeti, P. C., D. E. Reich, J. M. Higgins, H. Z. Levine, D. J. Richter *et al.*, 2002 Detecting recent positive selection in the human genome from haplotype structure. Nature 419: 832–837.

Saghai-Maroof, M. A., K. M. Soliman, R. A. Jorgensen, and R. W. Allard, 1984 Ribosomal DNA spacer-length polymorphisms in barley: mendelian inheritance, chromosomal location, and population dynamics. Proc. Natl. Acad. Sci. USA 81: 8014–8018.

Salvi, S., G. Sponza, M. Morgante, D. Tomes, X. Niu *et al.*, 2007 Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize. Proc. Natl. Acad. Sci. USA 104: 11376–11381.

Schmidt, R. J., F. A. Burr, M. J. Aukerman, and B. Burr, 1990 Maize regulatory gene opaque-2 encodes a protein with a "leucine-zipper" motif that binds to zein DNA. Proc. Natl. Acad. Sci. USA 87: 46–50.

Schmidt, R. J., M. Ketudat, M. J. Aukerman, and G. Hoschek, 1992 Opaque-2 is a transcriptional activator that recognizes a specific target site in 22-kD zein genes. Plant Cell 4: 689–700.

Schnable, P. S., D. Ware, R. S. Fulton, J. C. Stein, F. Wei *et al.*, 2009 The B73 maize genome: complexity, diversity, and dynamics. Science 326: 1112–1115.

Sekhon, R. S., H. Lin, K. L. Childs, C. N. Hansey, C. R. Buell *et al.*, 2011 Genome-wide atlas of transcription during maize development. Plant J. 66: 553–563.

Sekhon, R. S., C. N. Hirsch, K. L. Childs, M. W. Breitzman, P. Kell *et al.*, 2014 Phenotypic and transcriptional analysis of divergently selected maize populations reveals the role of developmental timing in seed size determination. Plant Physiol. 165: 658–669.

Springer, N. M., K. Ying, Y. Fu, T. Ji, C. T. Yeh *et al.*, 2009 Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. PLoS Genet. 5: e1000734.

Swanson-Wagner, R. A., S. R. Eichten, S. Kumari, P. Tiffin, J. C. Stein *et al.*, 2010 Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. Genome Res. 20: 1689–1699.

Tian, F., P. J. Bradbury, P. J. Brown, H. Hung, Q. Sun *et al.*, 2011 Genome-wide association study of leaf architecture in the maize nested association mapping population. Nat. Genet. 43: 159–162.

Turner, T. L., A. D. Stewart, A. T. Fields, W. R. Rice, and A. M. Tarone, 2011 Population-based resequencing of experimentally evolved populations reveals the genetic basis of body size variation in *Drosophila melanogaster*. PLoS Genet. 7: e1001336.

Venkatraman, E. S., and A. B. Olshen, 2007 A faster circular binary segmentation algorithm for the analysis of array CGH data. Bioinformatics 23: 657–663.

Waples, R. S., and C. Do, 2008 ldne: a program for estimating effective population size from data on linkage disequilibrium. Mol. Ecol. Resour. 8: 753–756.

Weir, B. S., and C. C. Cockerham, 1984 Estimating F-statistics for the analysis of population structure. Evolution 38: 1358–1370.

Wisser, R. J., S. C. Murray, J. M. Kolkman, H. Ceballos, and R. J. Nelson, 2008 Selection mapping of loci for quantitative disease resistance in a diverse maize population. Genetics 180: 583–599.

Wright, S., 1951 The genetical structure of populations. Ann. Eugen. 15: 323–354.

Yang, Z., E. J. van Oosterom, D. R. Jordan, A. Doherty, and G. L. Hammer, 2010 Genetic variation in potential kernel size affects kernel growth and yield of sorghum Crop Sci. 50: 685–695.

Yu, J., J. B. Holland, M. D. McMullen, and E. S. Buckler, 2008 Genetic design and statistical power of nested association mapping in maize. Genetics 178: 539–551.

Communicating editor: A. H. Paterson

GENETICS

Supporting Information
http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.167155/-/DC1

Insights into the Effects of Long-Term Artificial Selection on Seed Size in Maize

Candice N. Hirsch, Sherry A. Flint-Garcia, Timothy M. Beissinger, Steven R. Eichten, Shweta Deshpande, Kerrie Barry, Michael D. McMullen, James B. Holland, Edward S. Buckler, Nathan Springer, C. Robin Buell, Natalia de Leon, and Shawn M. Kaeppler

DOI: 10.1534/genetics.114.167155

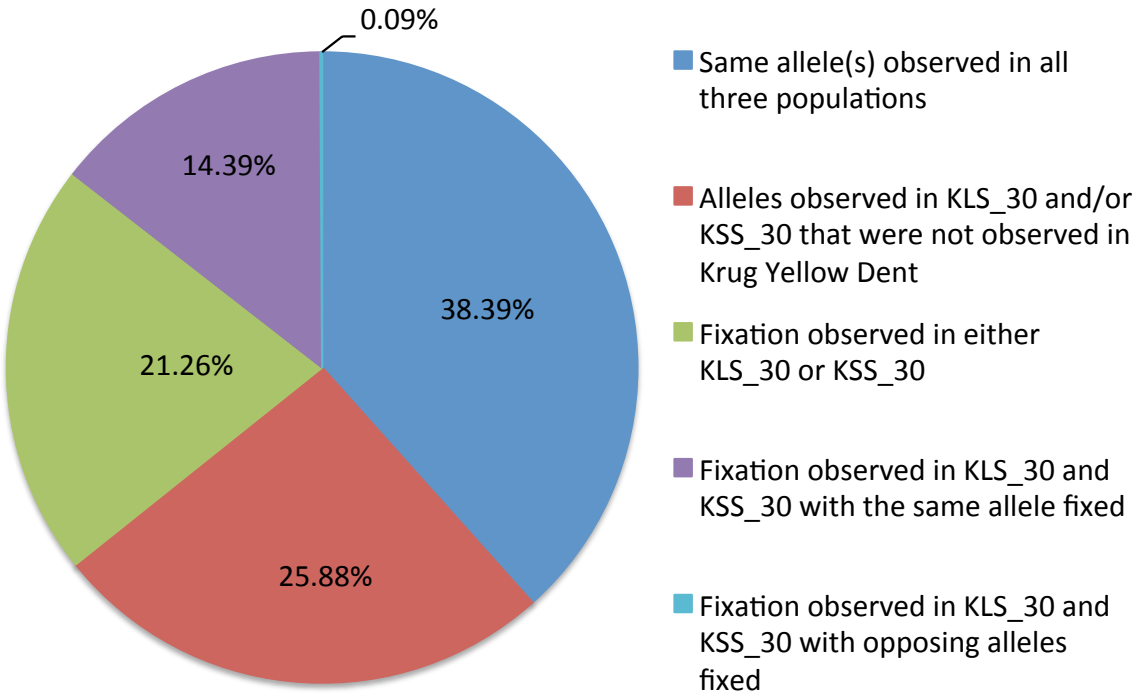
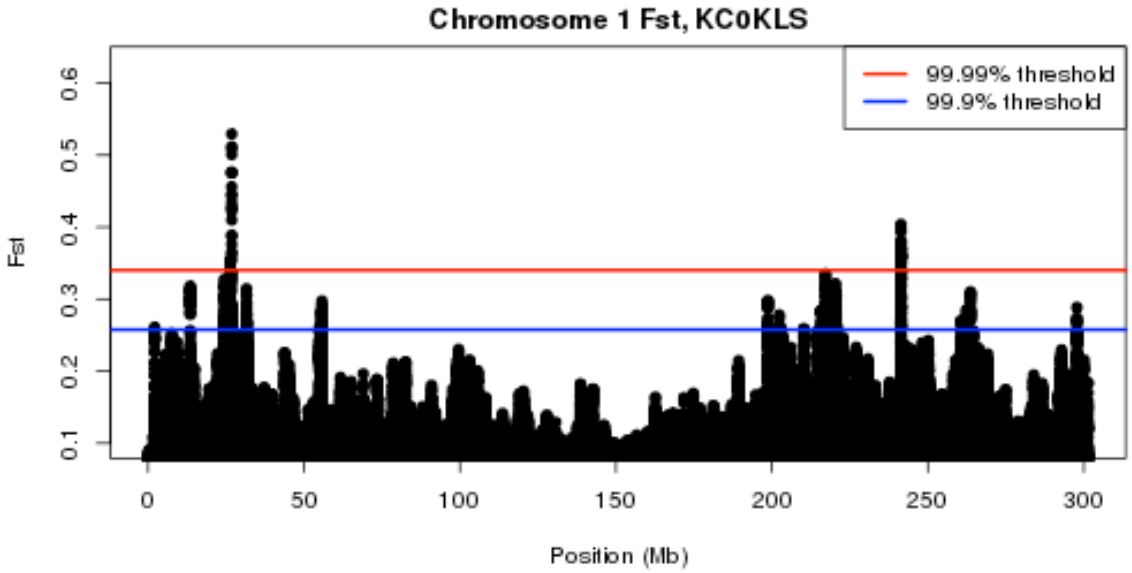
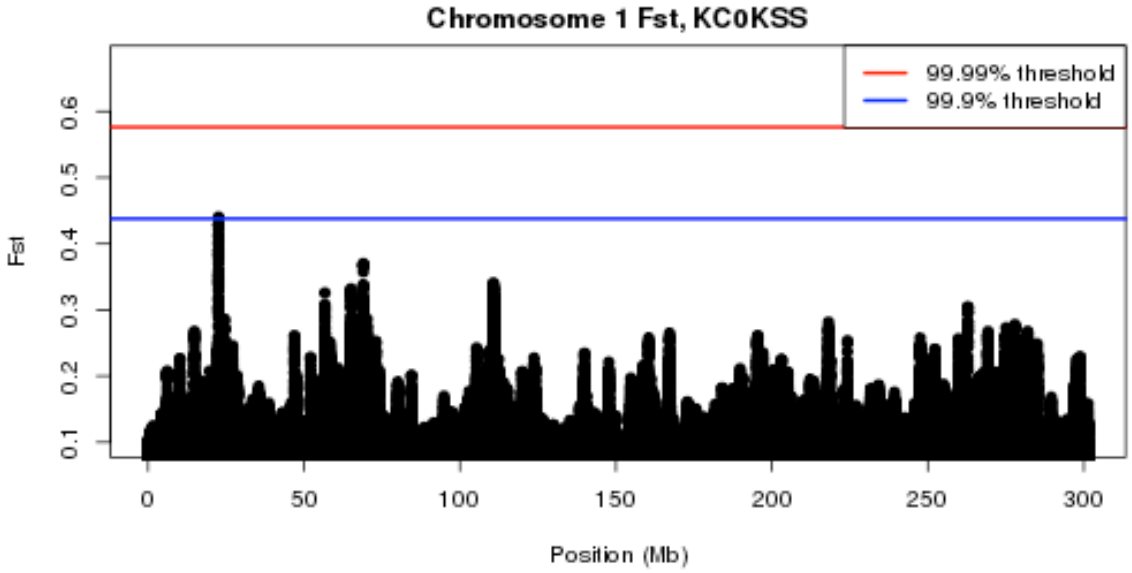
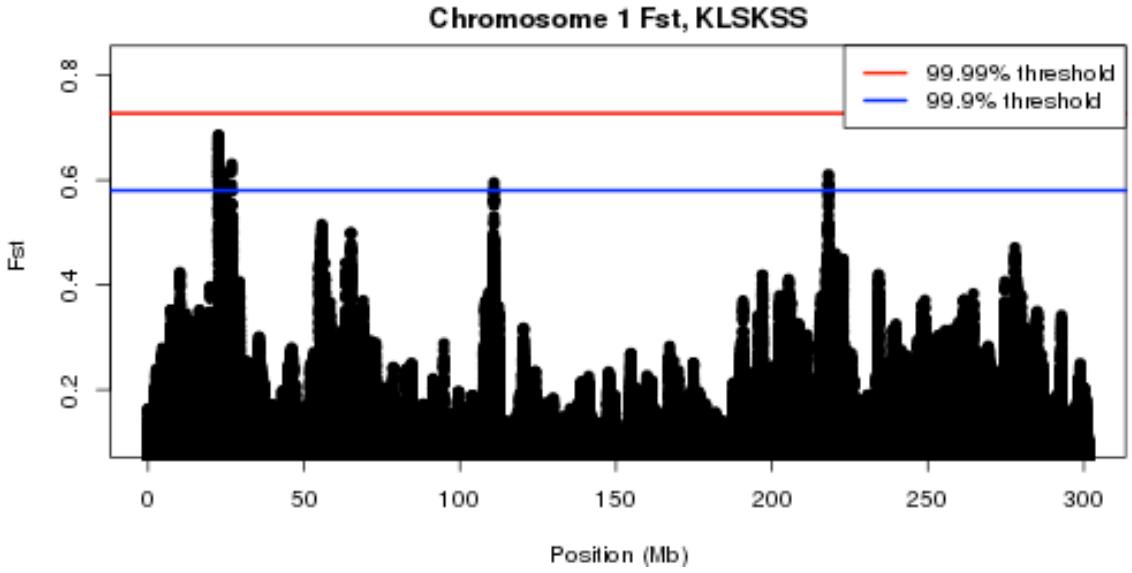
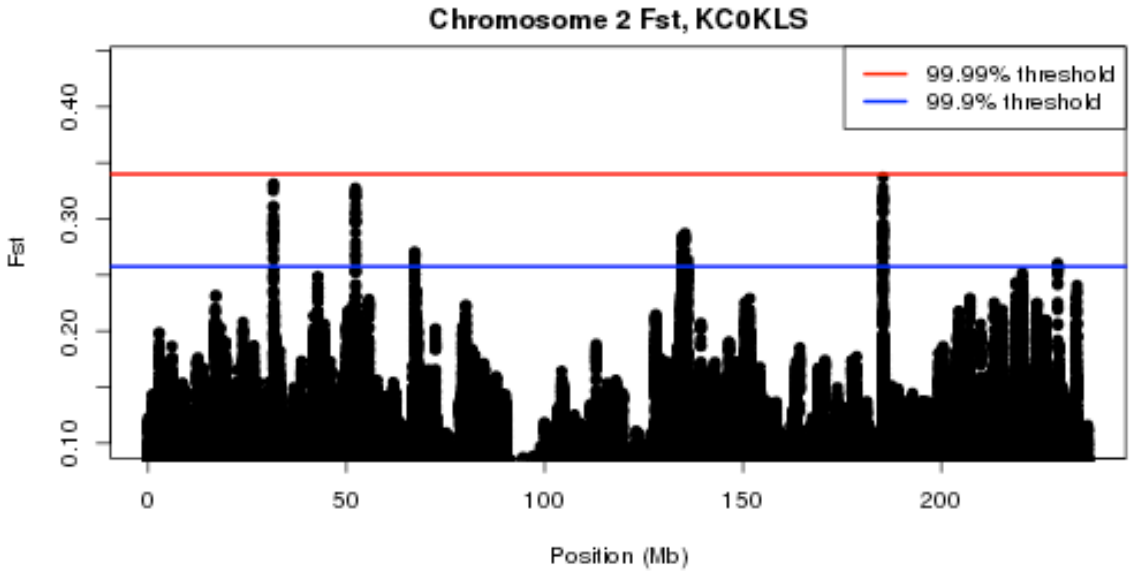


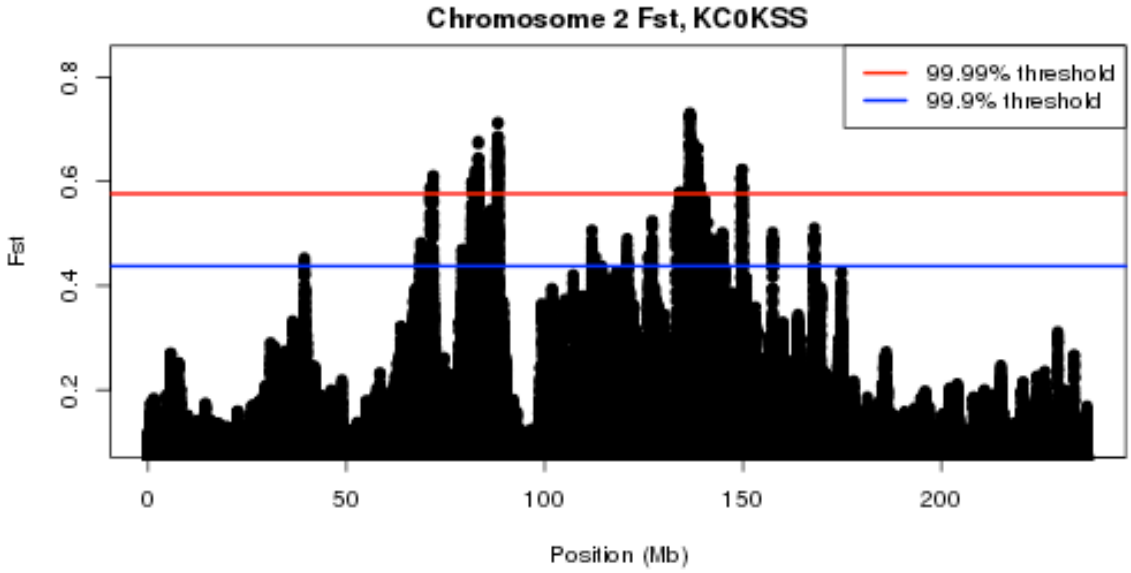
Figure S1 Categorization of single nucleotide polymorphism (SNP) variants within the populations Krug Yellow Dent, KLS_30, and KSS_30.

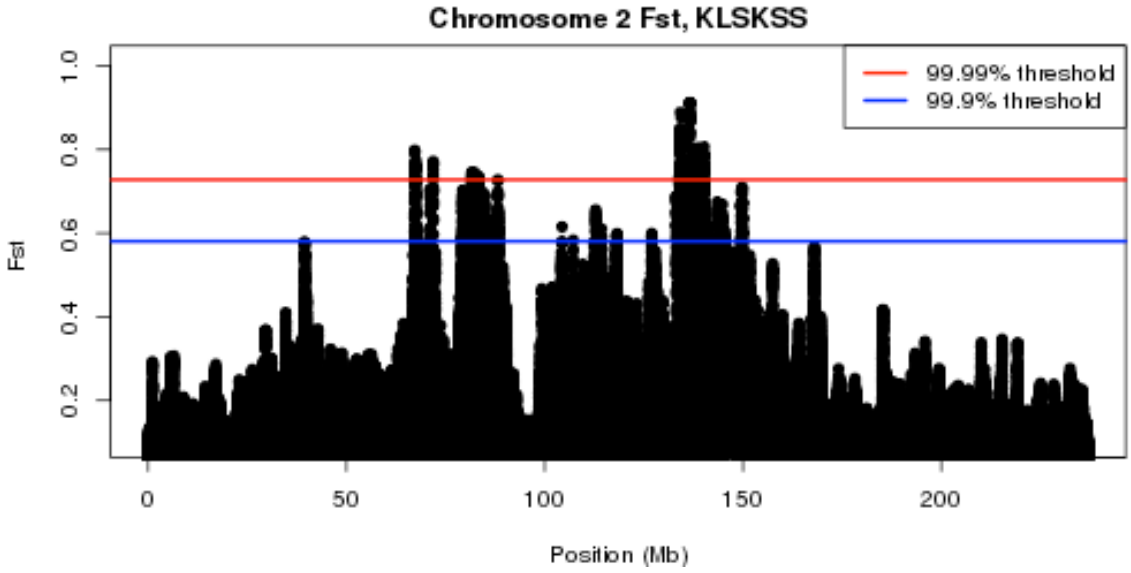

Chromosome 1 Fst, KC0KLS
99.99% threshold
99.9% threshold
Fst
0.1
0.2
0.3
0.4
0.5
0.6
0
50
100
150
200
250
300
Position (Mb)

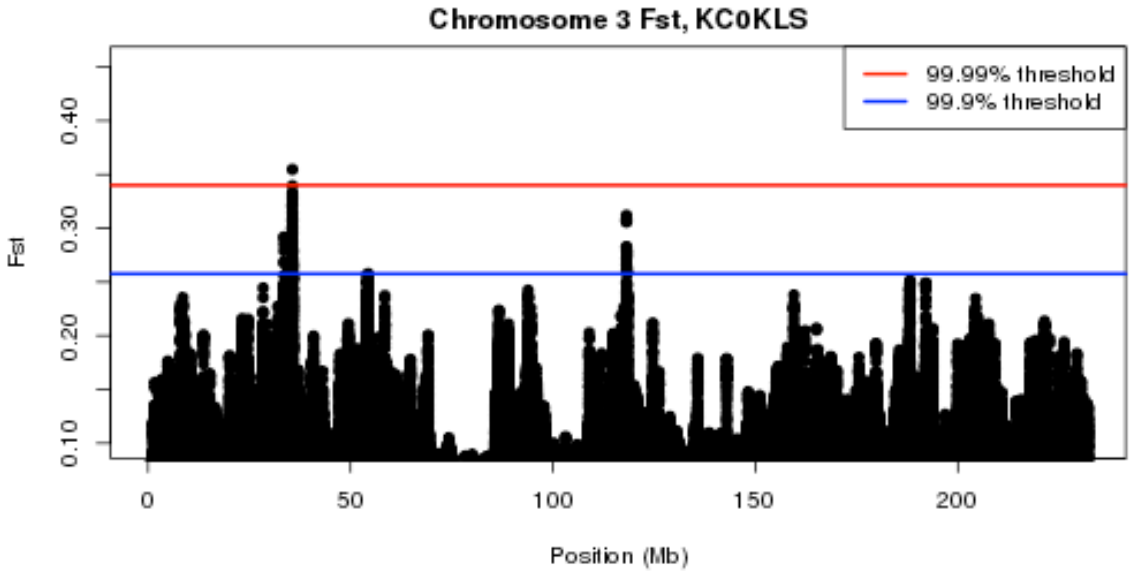

Chromosome 1 Fst, KC0KSS
99.99% threshold
99.9% threshold
Fst
0.1
0.2
0.3
0.4
0.5
0.6
0
50
100
150
200
250
300
Position (Mb)

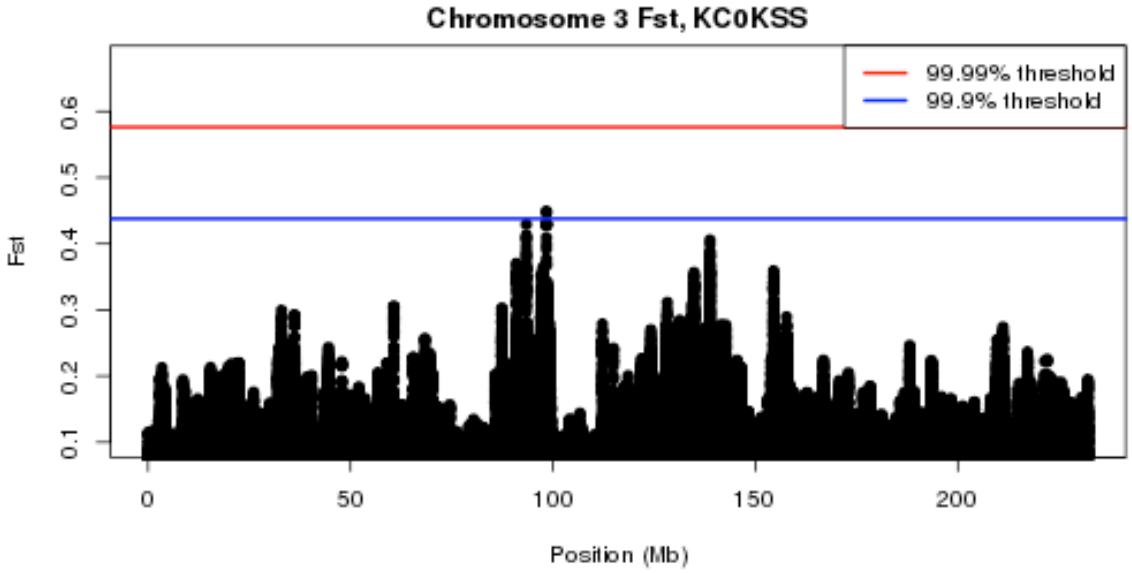

Chromosome 1 Fst, KLSKSS
99.99% threshold
99.9% threshold
Fst
0.2
0.4
0.6
0.8
0
50
100
150
200
250
300
Position (Mb)

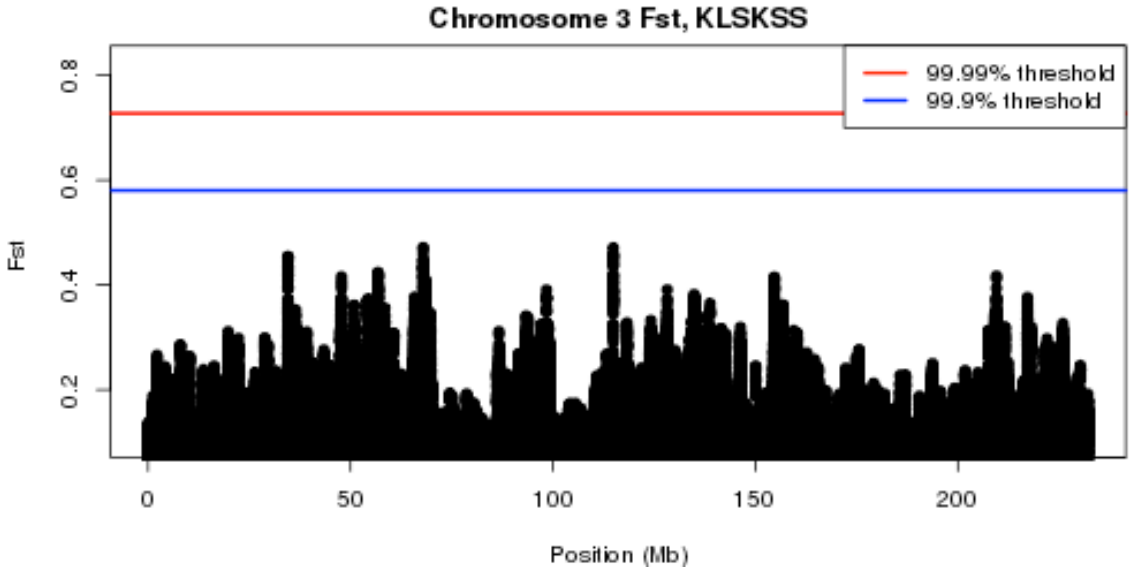

Chromosome 2 Fst, KC0KLS
99.99% threshold
99.9% threshold
Fst
0.10
0.20
0.30
0.40
0
50
100
150
200
Position (Mb)

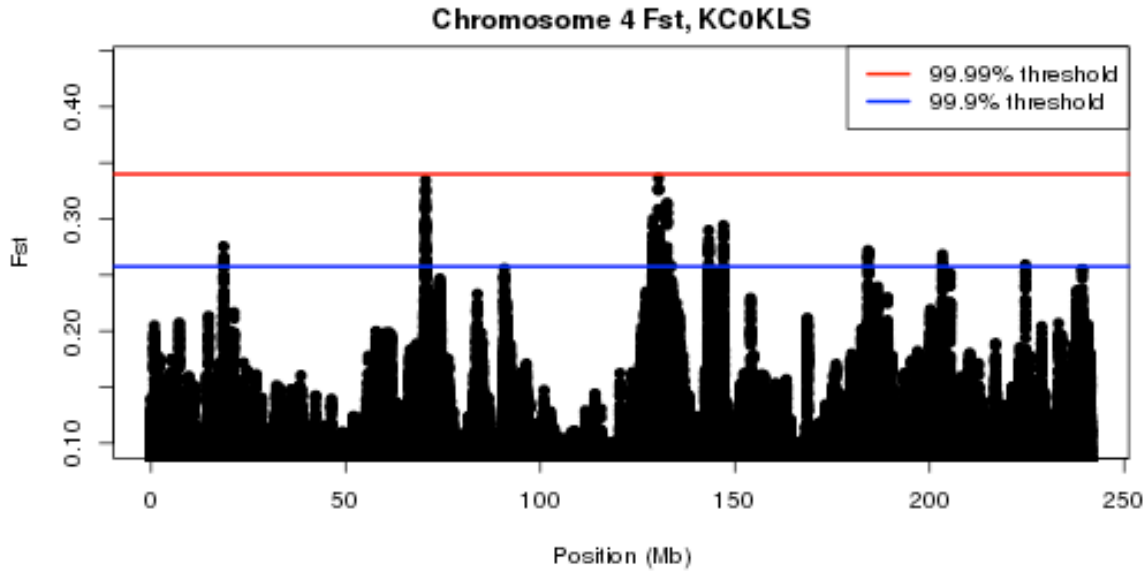

Chromosome 2 Fst, KC0KSS
99.99% threshold
99.9% threshold
Fst
0.2
0.4
0.6
0.8
0
50
100
150
200
Position (Mb)

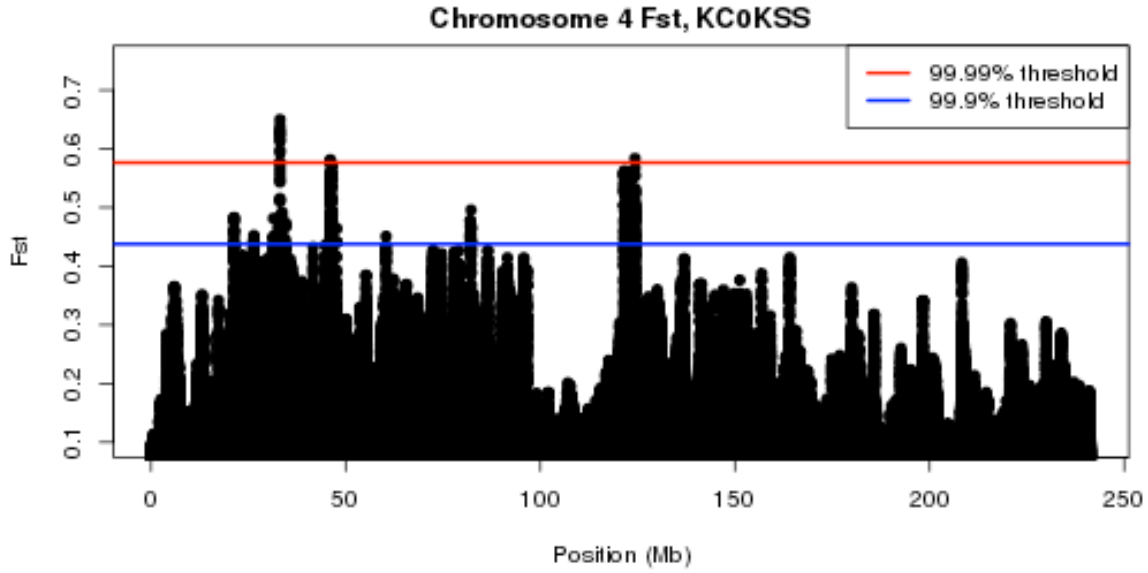

Chromosome 2 Fst, KLSKSS
99.99% threshold
99.9% threshold
Fst
0.2
0.4
0.6
0.8
1.0
0
50
100
150
200
Position (Mb)

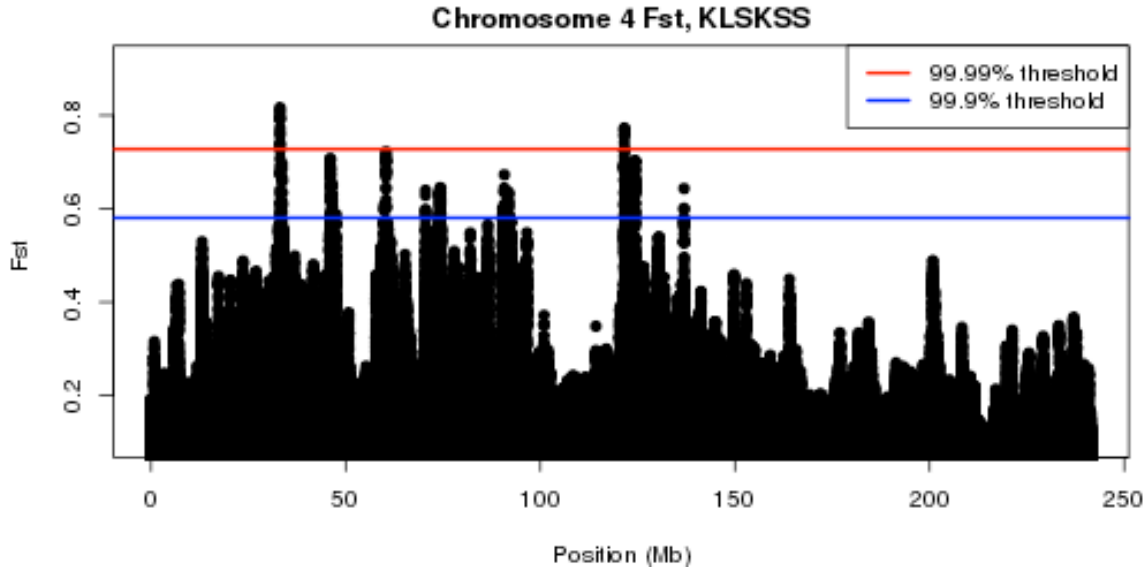

Chromosome 3 Fst, KC0KLS
99.99% threshold
99.9% threshold
Fst
0.10
0.20
0.30
0.40
0
50
100
150
200
Position (Mb)

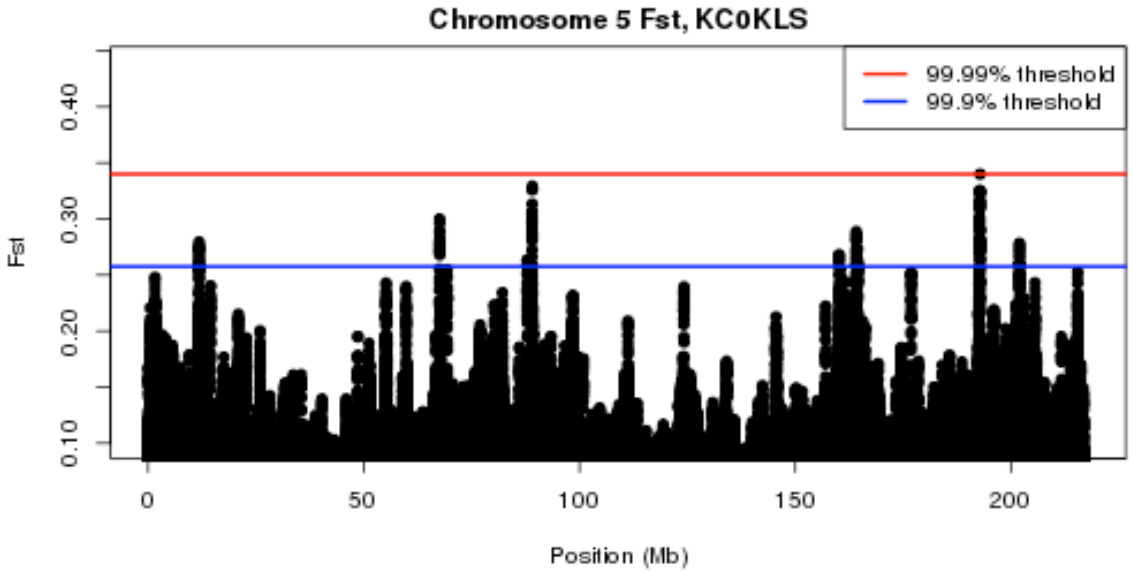

Chromosome 3 Fst, KC0KSS
99.99% threshold
99.9% threshold
Fst
0.1
0.2
0.3
0.4
0.5
0.6
0
50
100
150
200
Position (Mb)

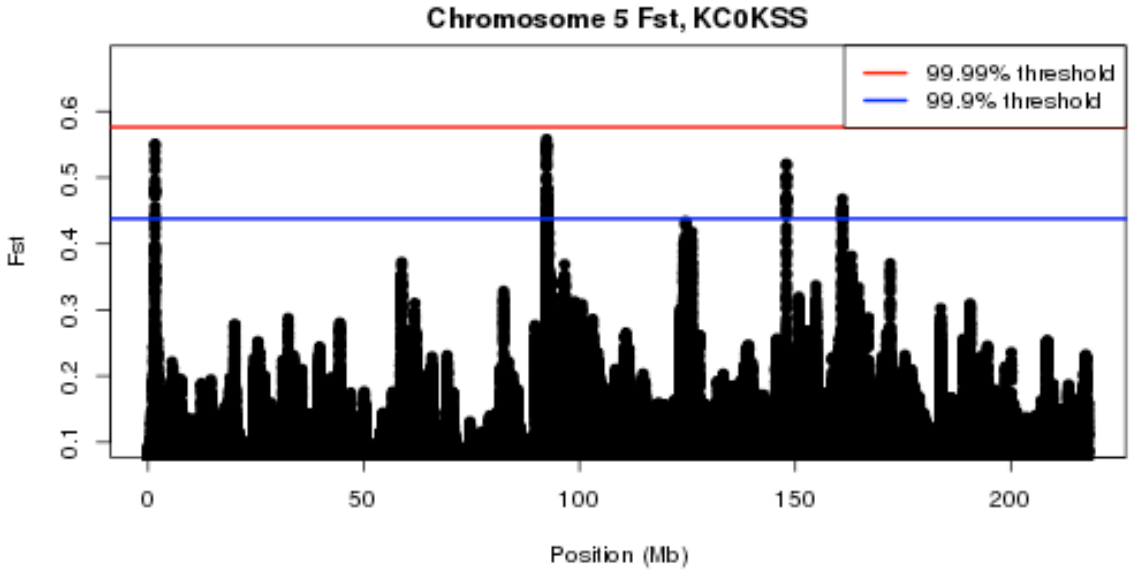

Chromosome 3 Fst, KLSKSS
99.99% threshold
99.9% threshold
Fst
0.2
0.4
0.6
0.8
0
50
100
150
200
Position (Mb)

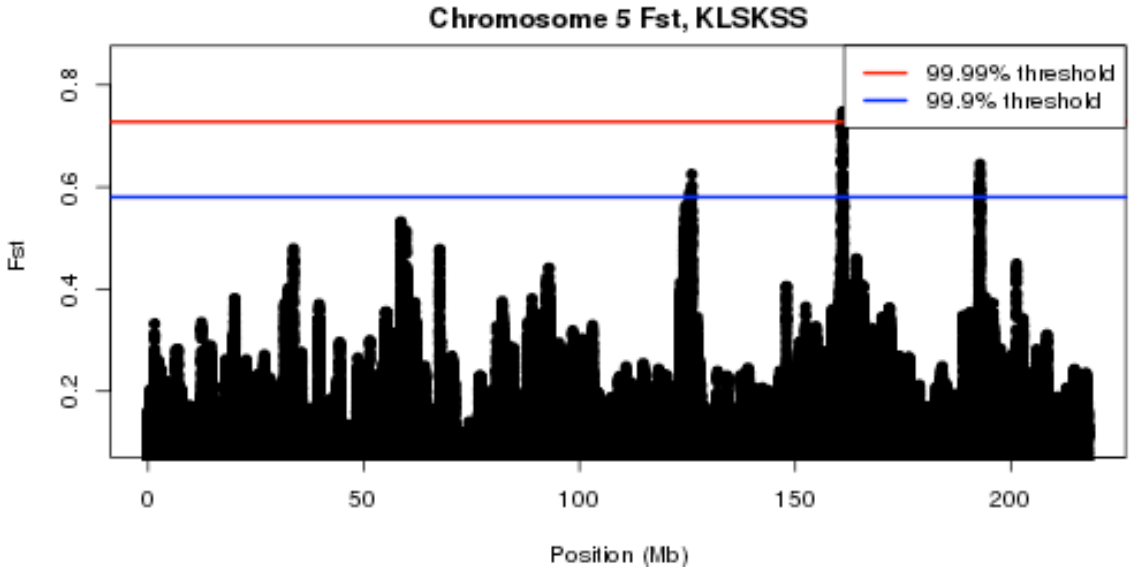

Chromosome 4 Fst, KC0KLS
99.99% threshold
99.9% threshold
Fst
0.10
0.20
0.30
0.40
0
50
100
150
200
250
Position (Mb)

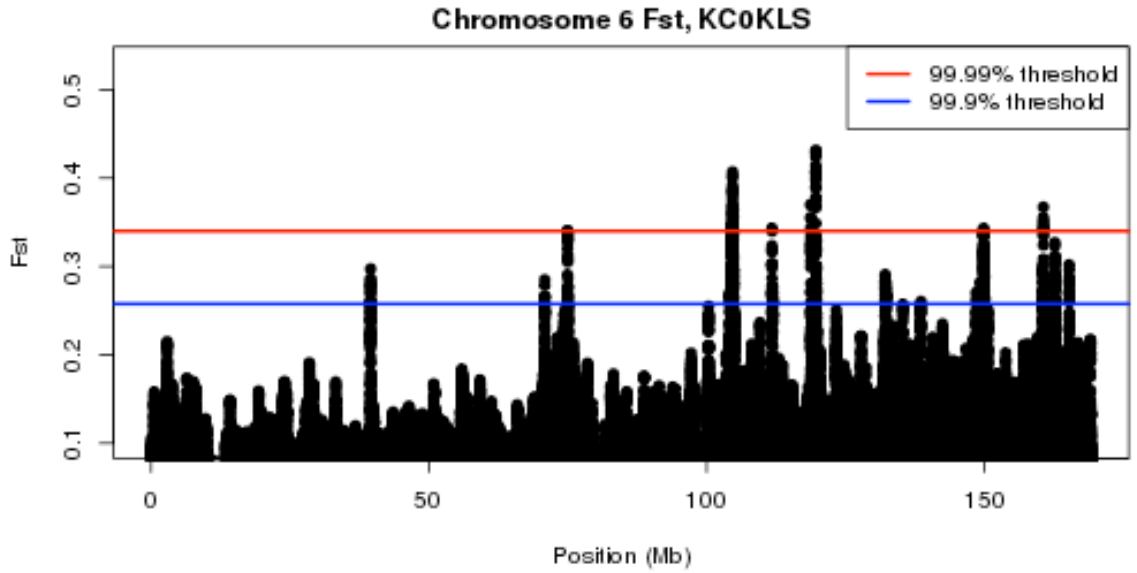

Chromosome 4 Fst, KC0KSS
99.99% threshold
99.9% threshold
Fst
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
50
100
150
200
250
Position (Mb)

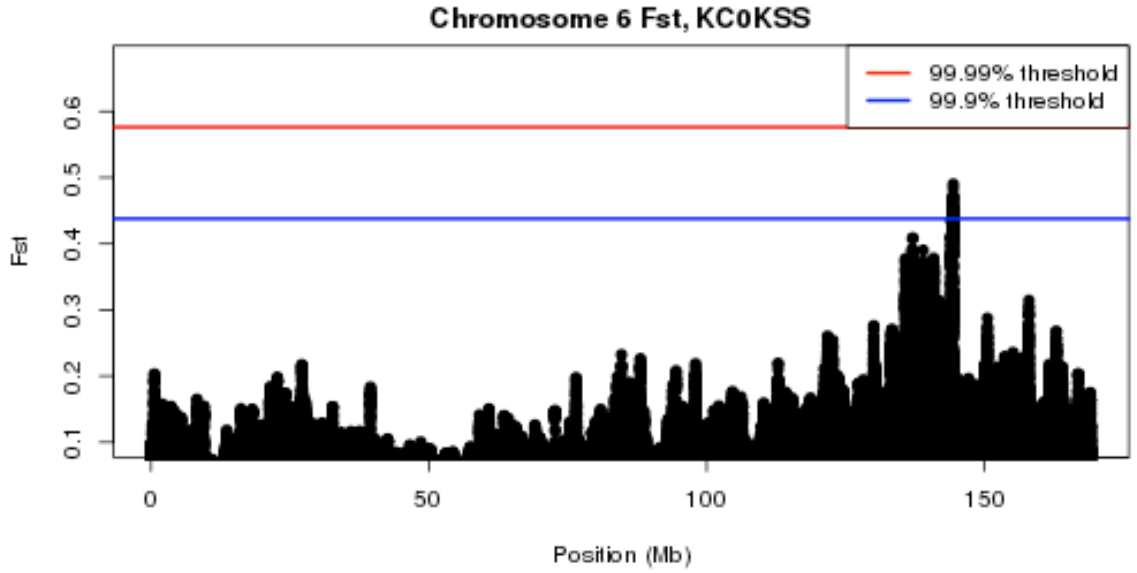

Chromosome 4 Fst, KLSKSS
99.99% threshold
99.9% threshold
Fst
0.2
0.4
0.6
0.8
0
50
100
150
200
250
Position (Mb)

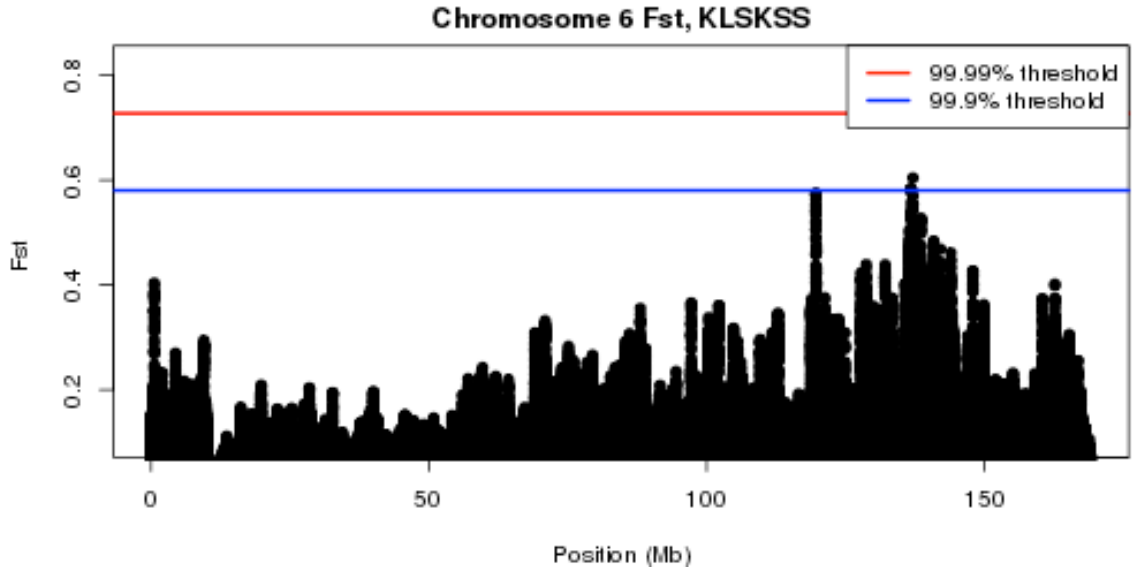

Chromosome 5 Fst, KC0KLS
99.99% threshold
99.9% threshold
Fst
0.10
0.20
0.30
0.40
0
50
100
150
200
Position (Mb)

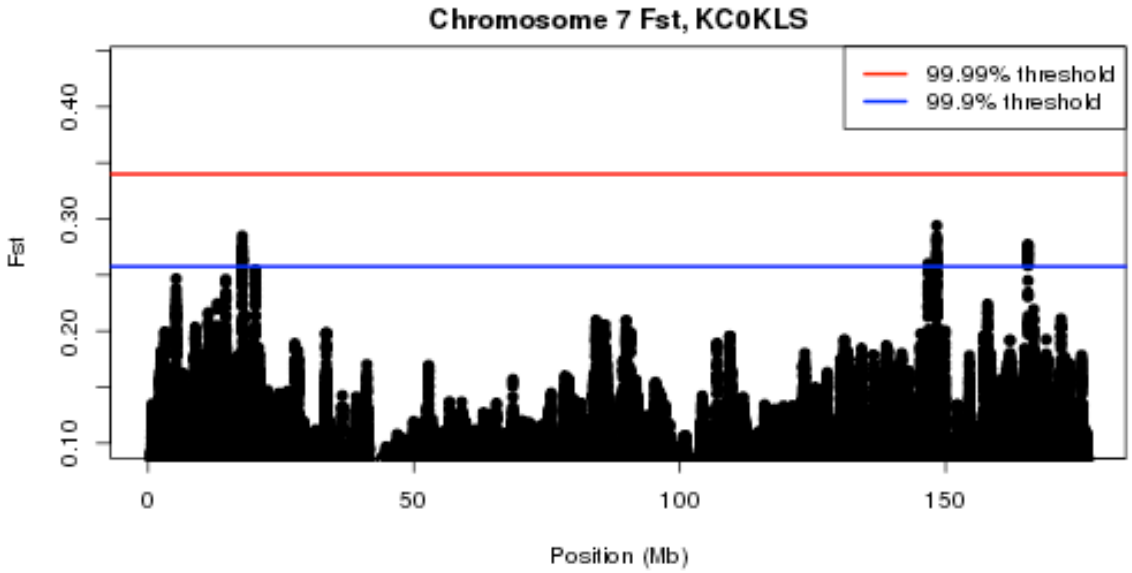

Chromosome 5 Fst, KC0KSS
99.99% threshold
99.9% threshold
Fst
0.1
0.2
0.3
0.4
0.5
0.6
0
50
100
150
200
Position (Mb)

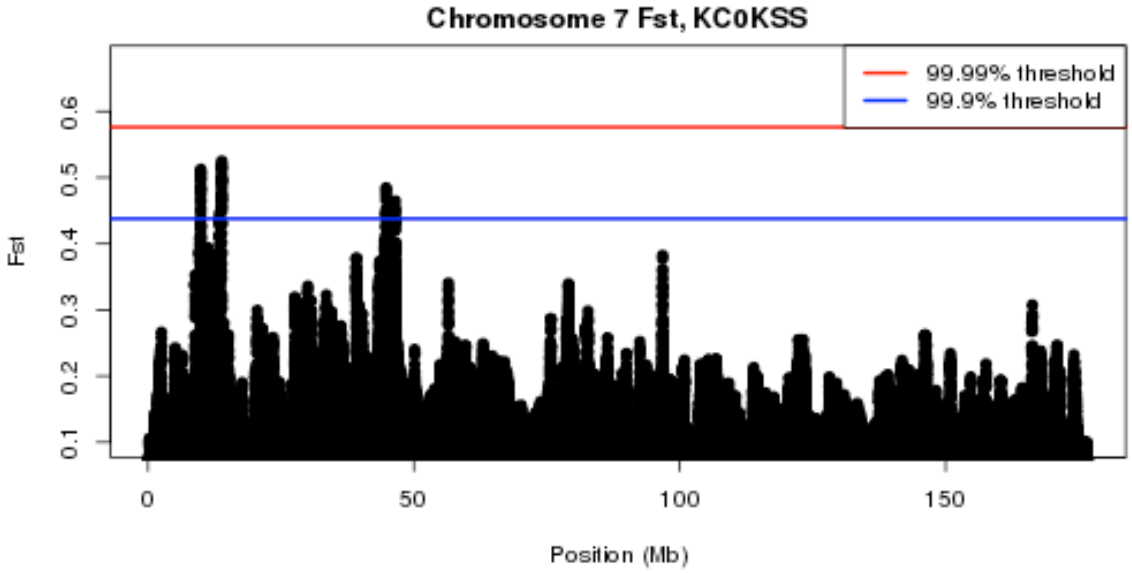

Chromosome 5 Fst, KLSKSS
99.99% threshold
99.9% threshold
Fst
0.2
0.4
0.6
0.8
0
50
100
150
200
Position (Mb)

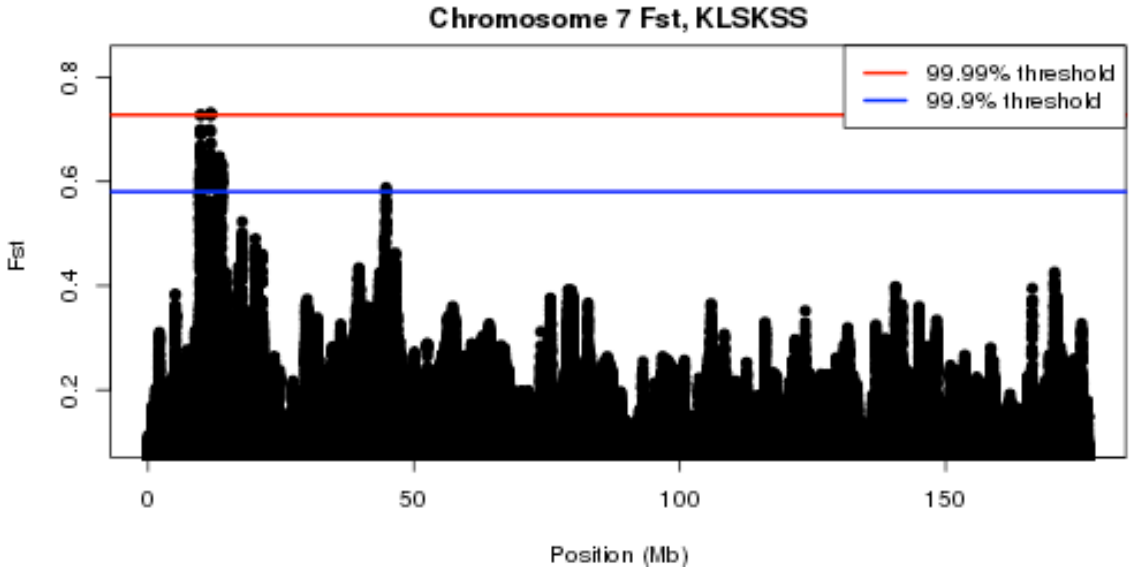

Chromosome 6 Fst, KC0KLS
99.99% threshold
99.9% threshold
Fst
0.1
0.2
0.3
0.4
0.5
0
50
100
150
Position (Mb)

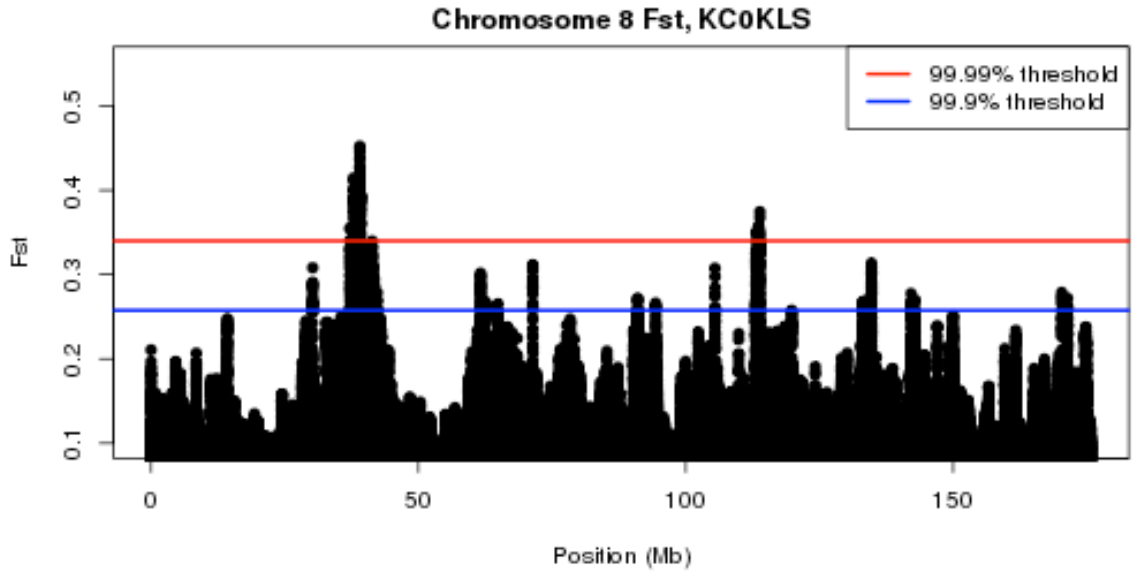

Chromosome 6 Fst, KC0KSS
99.99% threshold
99.9% threshold
Fst
0.1
0.2
0.3
0.4
0.5
0.6
0
50
100
150
Position (Mb)

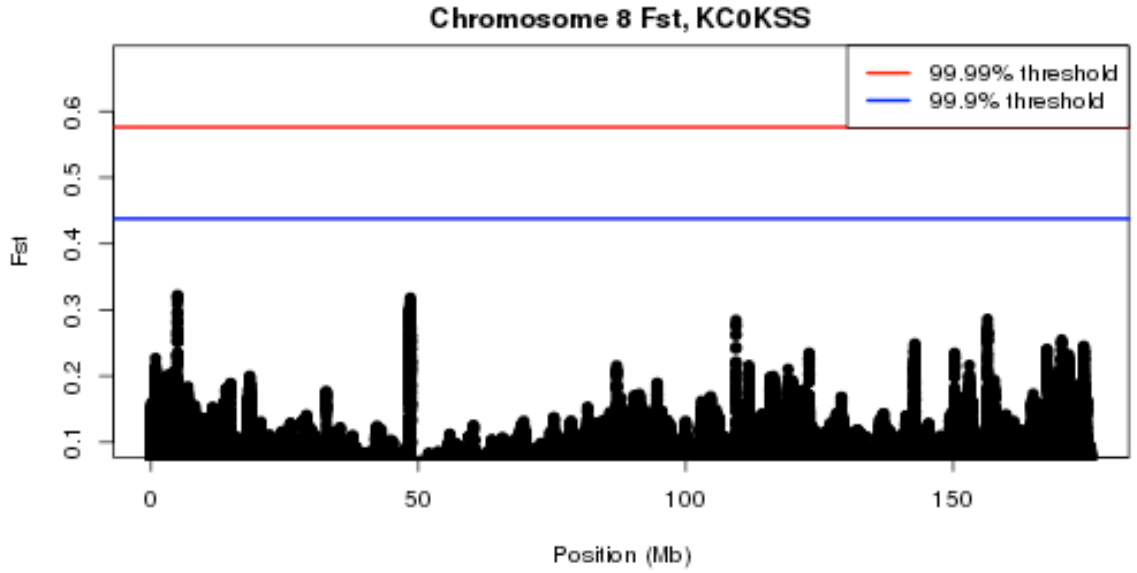

Chromosome 6 Fst, KLSKSS
99.99% threshold
99.9% threshold
Fst
0.2
0.4
0.6
0.8
0
50
100
150
Position (Mb)

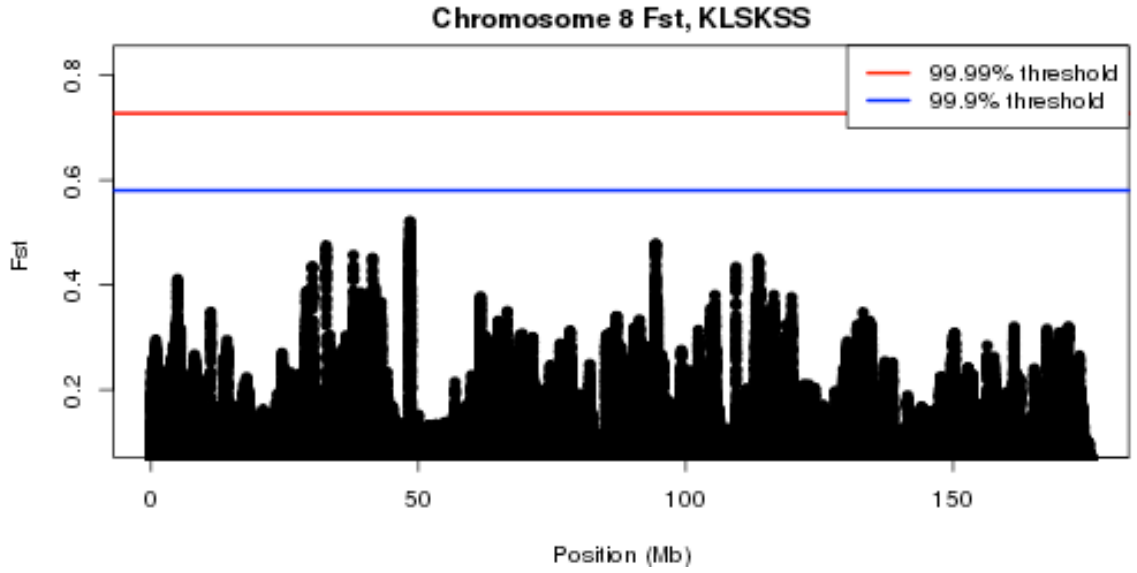

Chromosome 7 Fst, KC0KLS
99.99% threshold
99.9% threshold
Fst
0.10
0.20
0.30
0.40
0
50
100
150
Position (Mb)

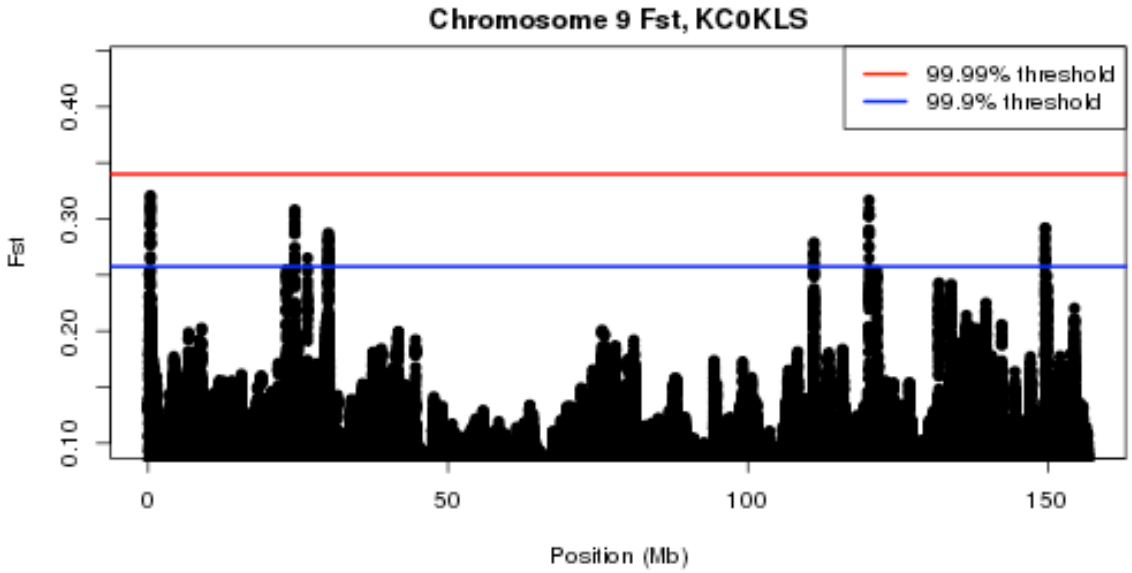

Chromosome 7 Fst, KC0KSS
99.99% threshold
99.9% threshold
Fst
0.1
0.2
0.3
0.4
0.5
0.6
0
50
100
150
Position (Mb)

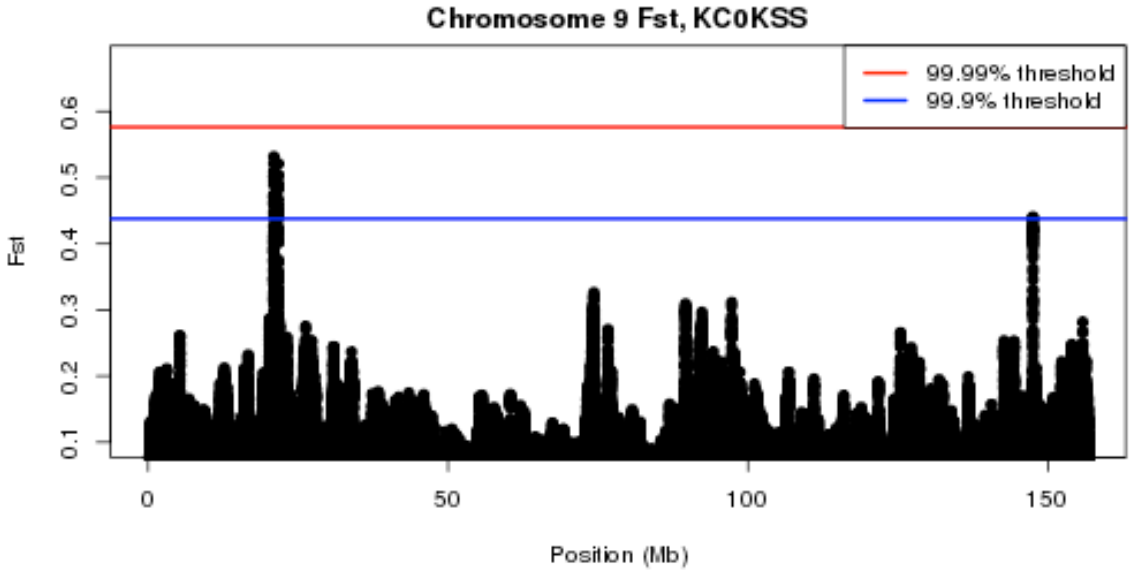

Chromosome 7 Fst, KLSKSS
99.99% threshold
99.9% threshold
Fst
0.2
0.4
0.6
0.8
0
50
100
150
Position (Mb)

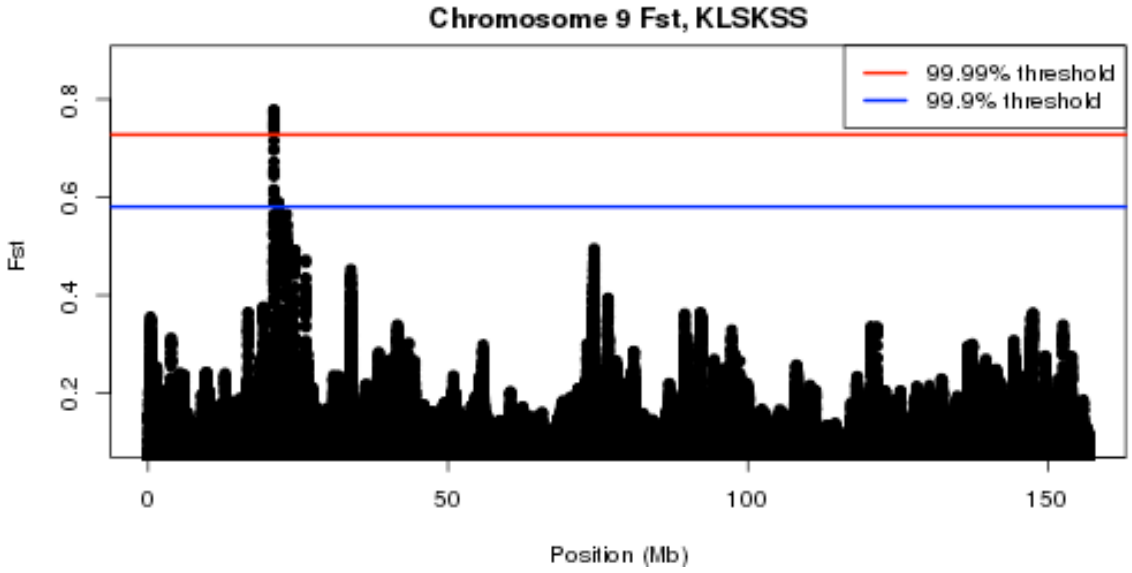

Chromosome 8 Fst, KC0KLS
99.99% threshold
99.9% threshold
Fst
0.1
0.2
0.3
0.4
0.5
0
50
100
150
Position (Mb)

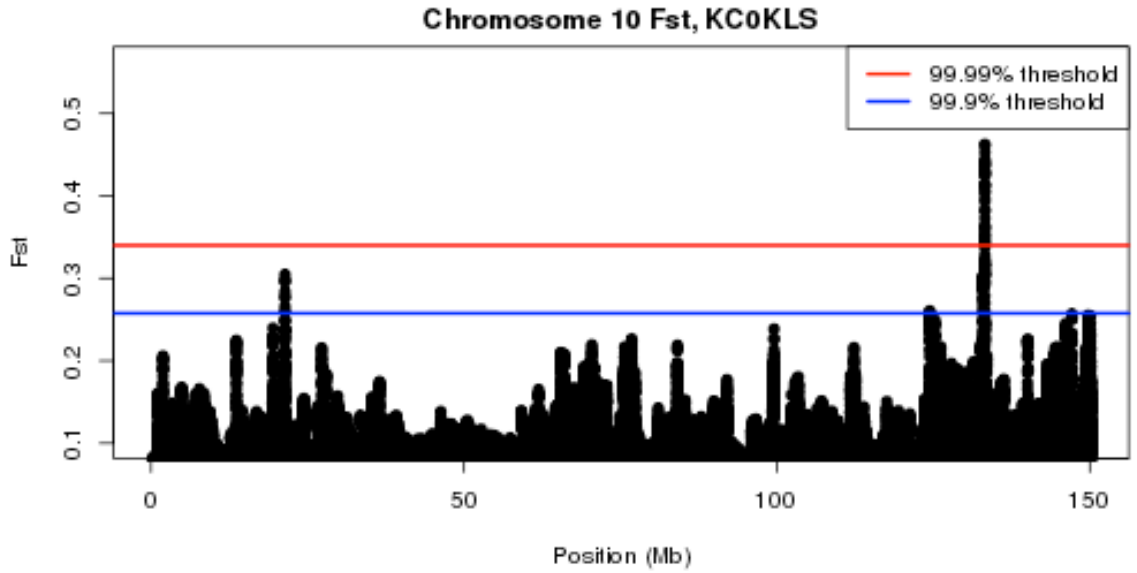

Chromosome 8 Fst, KC0KSS
99.99% threshold
99.9% threshold
Fst
0.1
0.2
0.3
0.4
0.5
0.6
0
50
100
150
Position (Mb)

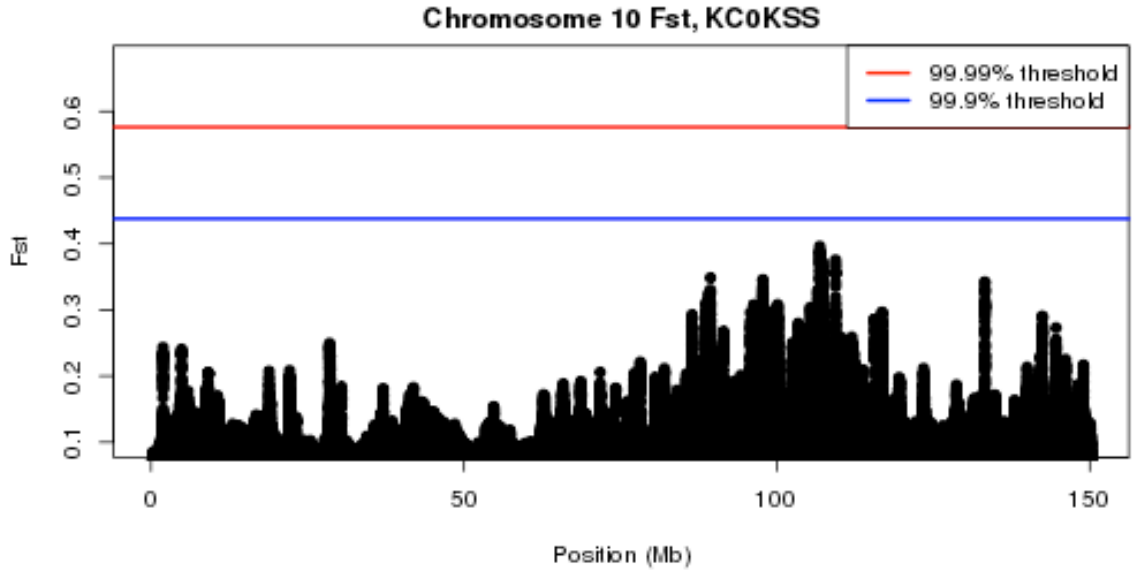

Chromosome 8 Fst, KLSKSS
99.99% threshold
99.9% threshold
Fst
0.2
0.4
0.6
0.8
0
50
100
150
Position (Mb)


Chromosome 9 Fst, KC0KLS
99.99% threshold
99.9% threshold
Fst
0.10
0.20
0.30
0.40
0
50
100
150
Position (Mb)


Chromosome 9 Fst, KC0KSS
99.99% threshold
99.9% threshold
Fst
0.1
0.2
0.3
0.4
0.5
0.6
0
50
100
150
Position (Mb)


Chromosome 9 Fst, KLSKSS
99.99% threshold
99.9% threshold
Fst
0.2
0.4
0.6
0.8
0
50
100
150
Position (Mb)


Chromosome 10 Fst, KC0KLS
99.99% threshold
99.9% threshold
Fst
0.1
0.2
0.3
0.4
0.5
0
50
100
150
Position (Mb)


Chromosome 10 Fst, KC0KSS
99.99% threshold
99.9% threshold
Fst
0.1
0.2
0.3
0.4
0.5
0.6
0
50
100
150
Position (Mb)

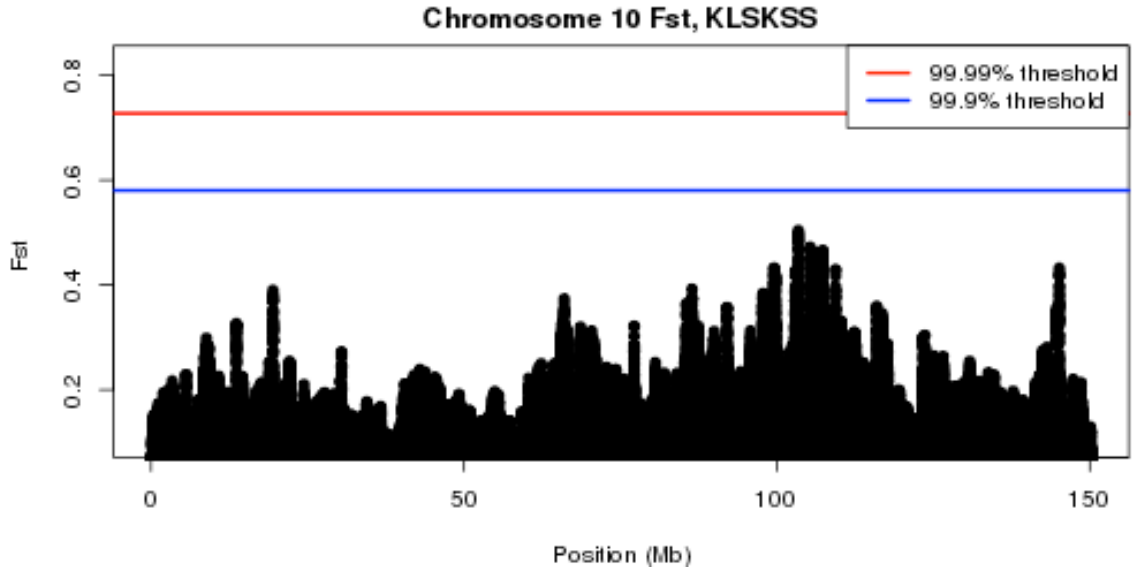

Chromosome 10 Fst, KLSKSS
99.99% threshold
99.9% threshold
Fst
0.2
0.4
0.6
0.8
0
50
100
150
Position (Mb)

Figure S2 F_{ST} values for each of the maize chromosomes. F_{ST} values were calculated using a 25-single nucleotide polymorphism (SNP) sliding window approach. Comparisons were made between Krug Yellow Dent and KLS_30, Krug Yellow Dent and KSS_30, and KLS_30 and KSS_30.

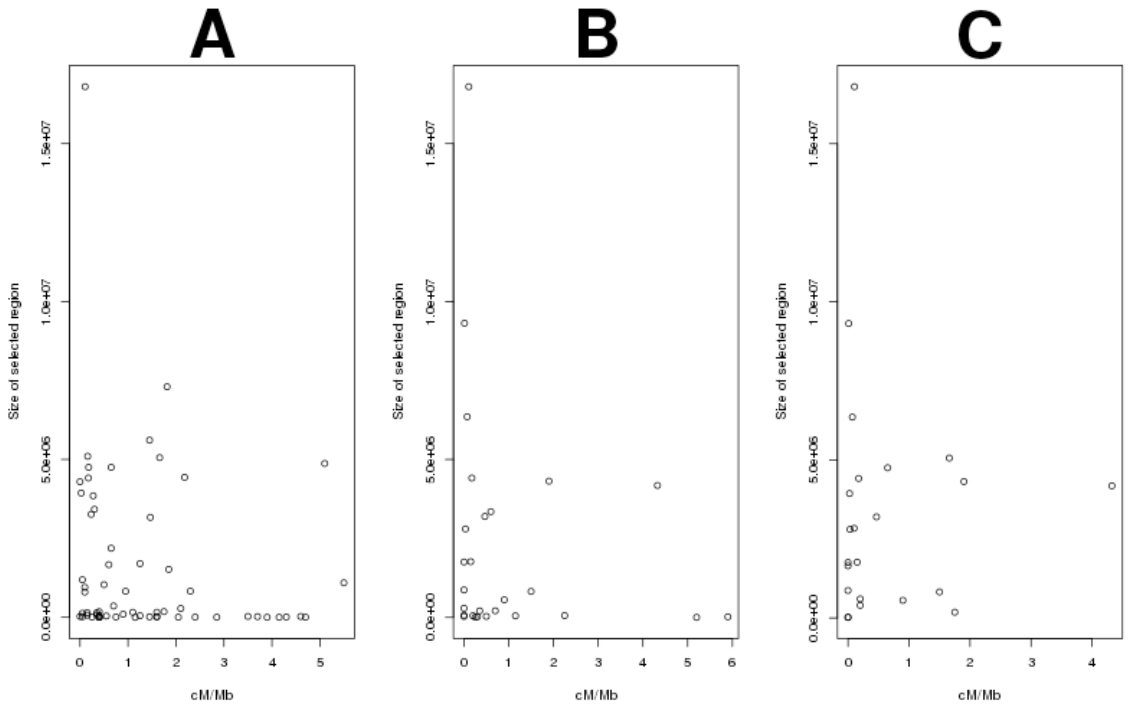


Figure S3 Region size versus relative recombination rate for each region identified as putatively under selection in the Krug long-term selection populations at the 99.9% outlier threshold. A) Regions identified by comparing Krug Yellow Dent to KLS_30, B) Regions identified by comparing Krug Yellow Dent to KSS_30, C) Regions identified by comparing KLS_30 and KSS_30. For all, relative levels of recombination across the genome were approximated based on recombination frequencies in the intermated B73 x Mo17 population. No significant correlations were observed.

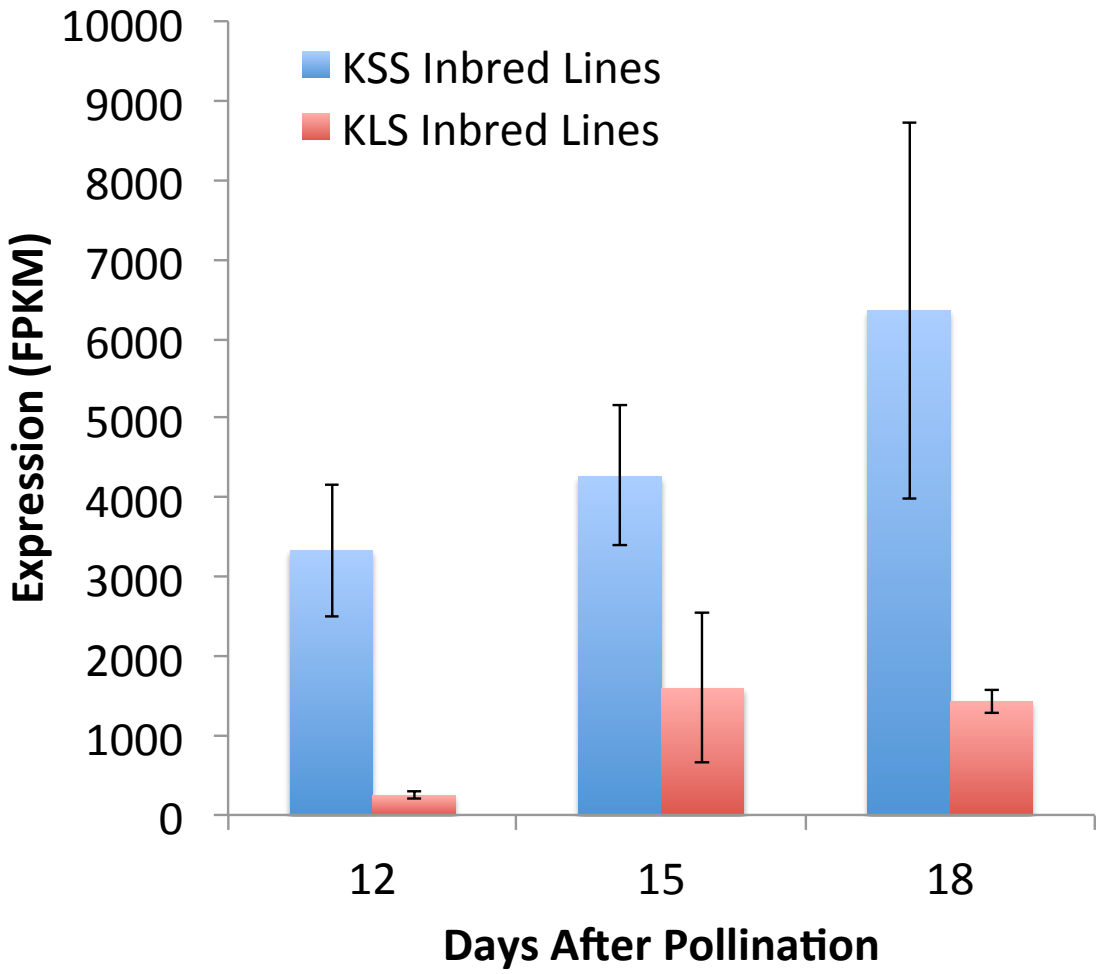


Figure S4 Average endosperm transcript abundance estimates for inbred lines derived from the KSS_30 and KLS_30 populations for the *Opaque2* gene. Error bars show standard deviations calculated from three biological replicates. Data for this figure was obtained from (SEKHON *et al.* 2014).

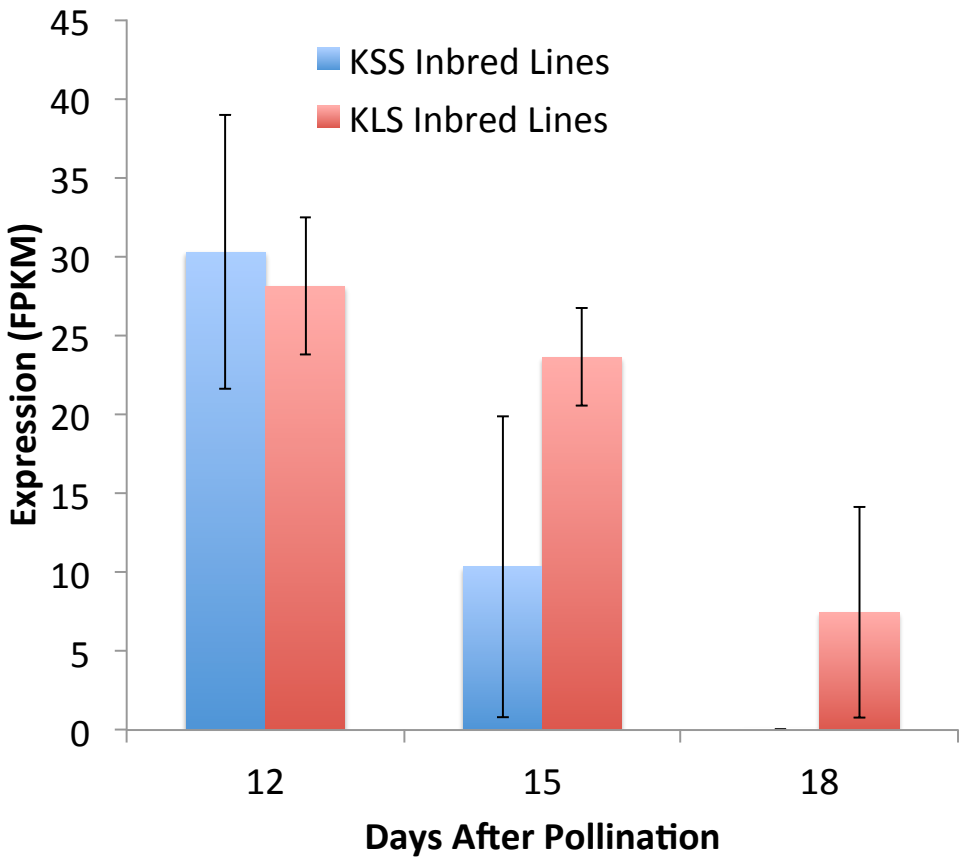


Figure S5 Average endosperm transcript abundance estimates for inbred lines derived from the KSS_30 and KLS_30 populations for the gene GRMZM2G069078. Error bars show standard deviations calculated from three biological replicates. Data for this figure was obtained from (SEKHON *et al.* 2014)

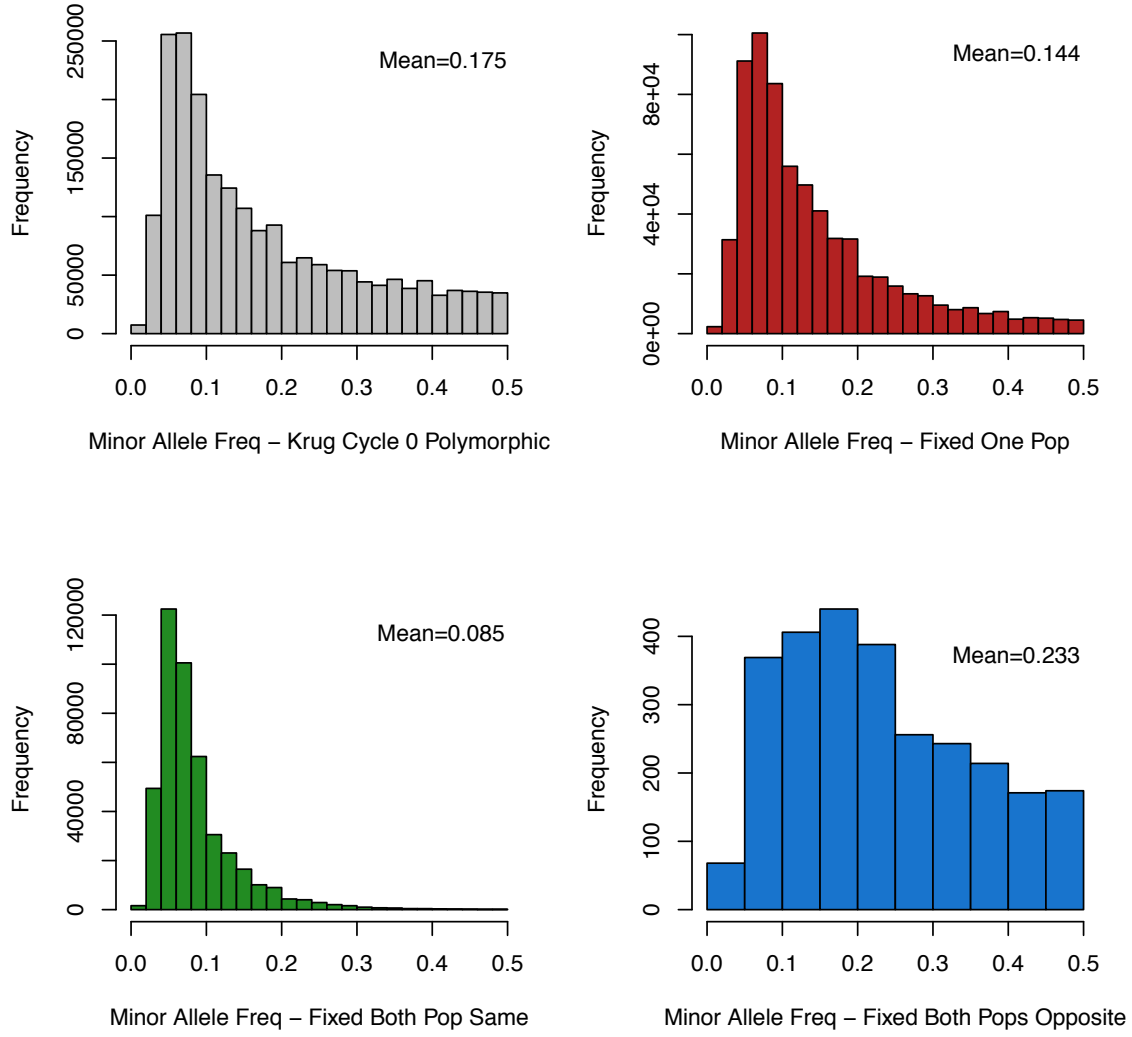
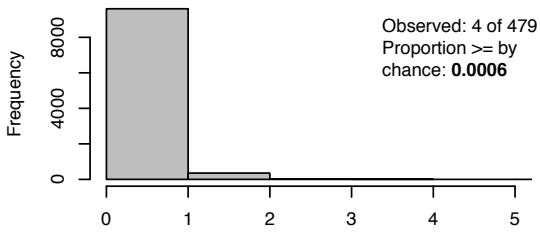
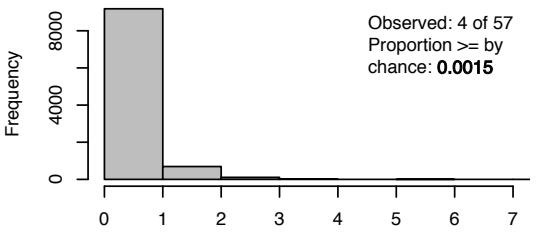


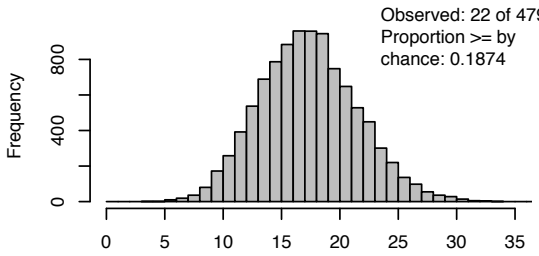
Figure S6 Empirical minor allele frequency for 2,056,663 SNPs that were polymorphic in the Krug Yellow Dent population and subsets of these SNPs that were fixed in one or both of the selected populations. 664,056 SNPs reached fixation in only one population (red), 444,599 SNPs reached fixation in both populations with the same fixed allele (green), and 2,729 SNPs reads in both populations reached fixation in both populations with oppositely fixed SNPs (blue).



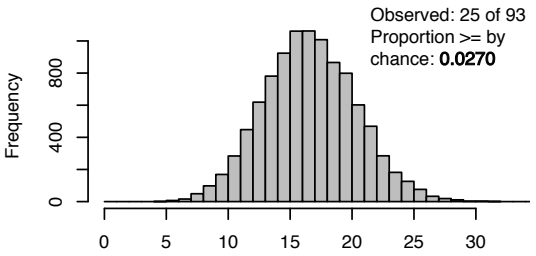
Overlap Random CGH Regions with Observed SeqCNV Regions



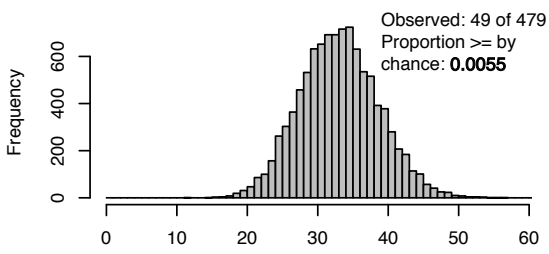
Overlap Random SeqCNV Regions with Observed CGH Regions



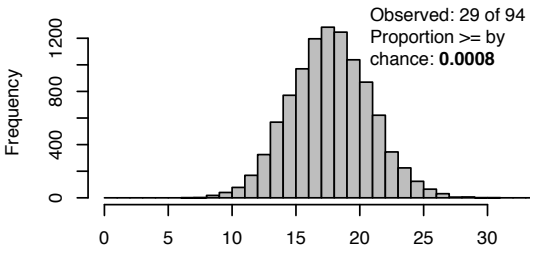
Overlap Random CGH Regions with Observed NAM Regions



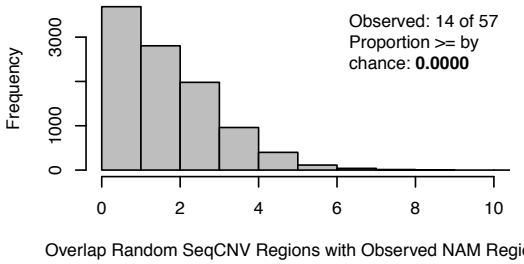
Overlap Random NAM Regions with Observed CGH Regions



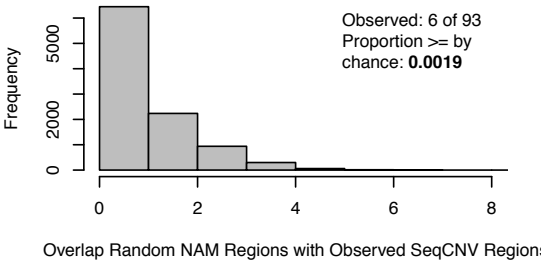
Overlap Random CGH Regions with Observed Sweep Regions



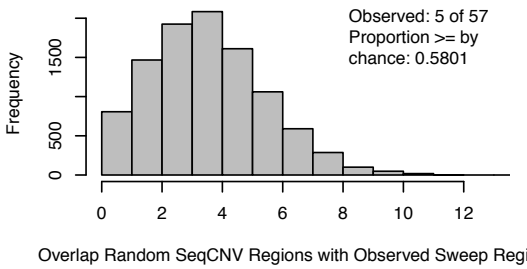
Overlap Random Sweep Regions with Observed CGH Regions



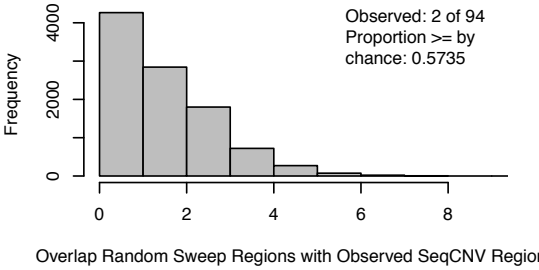
Overlap Random SeqCNV Regions with Observed NAM Regions



Overlap Random NAM Regions with Observed SeqCNV Regions



Overlap Random SeqCNV Regions with Observed Sweep Regions



Overlap Random Sweep Regions with Observed SeqCNV Regions

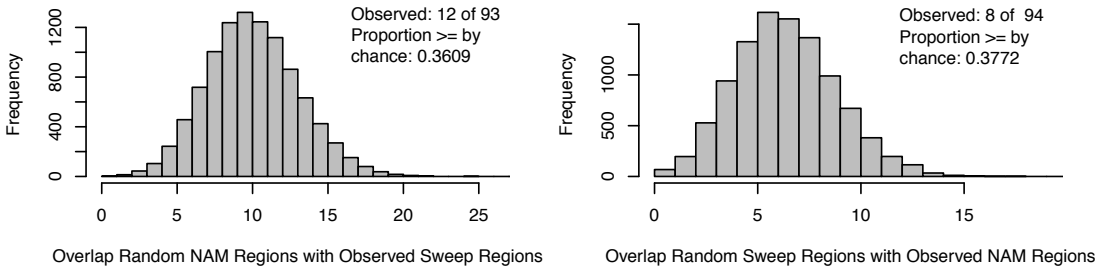


Figure S7 Simulation experiment testing the pair-wise overlap between each source of evidence [comparative genome hybridization (CGH) copy number variation (CNV) regions, sequence depth CNV regions (SeqCNV), regions exceeding the 99.9% outlier threshold (Sweep Regions), and regions identified in the nested association mapping (NAM) population] by chance compared with the empirically observed overlap. Regions with the empirically observed size were randomly placed throughout the genome 10,000 for each source of evidence. Comparisons were then made between the random data and observed data to test the overlap that was observed by chance.

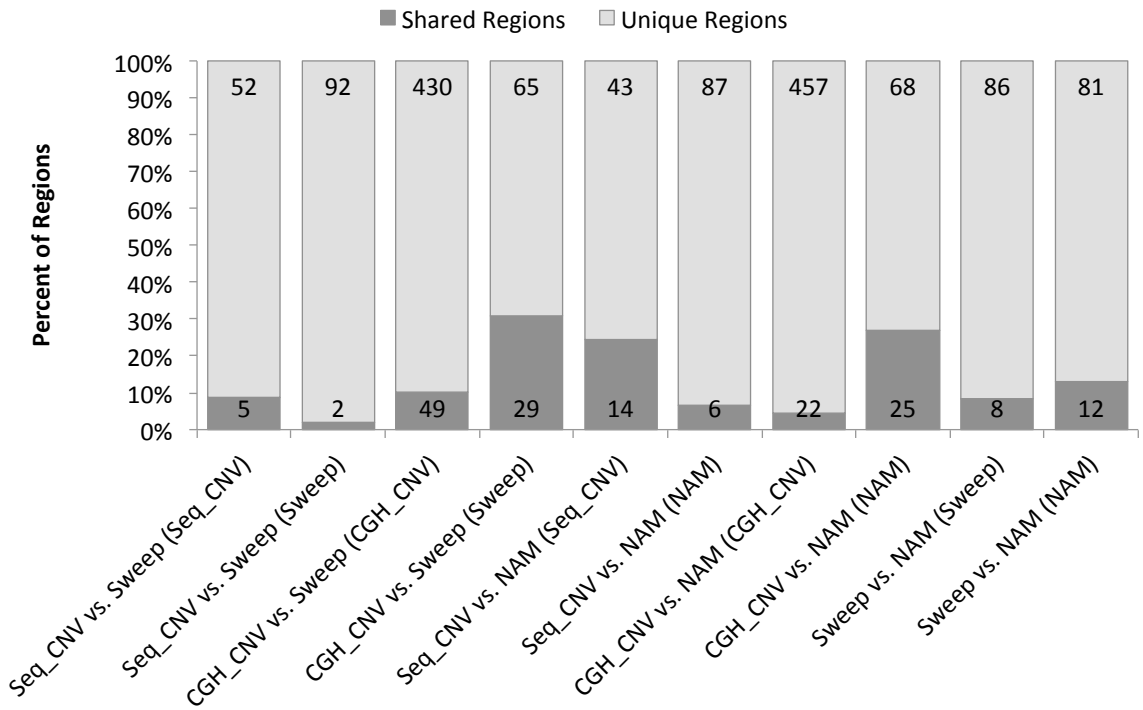


Figure S8 Pair-wise comparisons of overlapping variable regions in the Krug Yellow Dent divergent long-term selection experiment for seed size and quantitative trait loci for seed weight in the maize nested association (NAM) population. Type of variation in parenthesis following the pair-wise comparison description indicates which type of variation the bar pertains to. A comparison with NAM SNPs required regions to be within 500kb to be considered shared regions and for all other comparisons 10kb overlap was required.

Table S1 Candidate regions under selection during 30 generations of selection for seed size, at the 99.9% level. Regions were identified using F_{ST} values and a 25-single nucleotide polymorphism (SNP) sliding window approach. Comparisons were made between Krug Yellow Dent and KLS_30, Krug Yellow Dent and KSS_30, and KLS_30 and KSS_30. One indicates a difference and zero indicates no difference for the region.

Region	Chromosome	Start	End	Krug Yellow Dent vs. KLS_30	Krug Yellow Dent vs. KSS_30	KLS_30 vs. KSS_30	Number of Genes in Region
1	chr1	2088034	2099502	1	0	0	2
2	chr1	13507647	13537594	1	0	0	1
3	chr1	22511847	26826200	0	1	1	103
4	chr1	24438690	31739806	1	0	0	169
5	chr1	54928624	55758694	1	0	0	18
6	chr1	110877561	110898319	0	0	1	0
7	chr1	198652274	203086089	1	0	0	100
8	chr1	210227884	210245719	1	0	0	1
9	chr1	215729004	220787144	1	0	1	92
10	chr1	241280862	241441394	1	0	0	7
11	chr1	260539402	263706680	1	0	0	74
12	chr1	297792787	297796118	1	0	0	1
13	chr10	21394962	21401418	1	0	0	1
14	chr10	124383915	124428189	1	0	0	1
15	chr10	132885971	133247775	1	0	0	9
16	chr2	31592504	31644896	1	0	0	2
17	chr2	39419610	39427387	0	1	0	2
18	chr2	52315602	52324852	1	0	0	1
19	chr2	67149060	71899682	1	0	1	44
20	chr2	68731886	72080518	0	1	0	33
21	chr2	79033475	88346653	0	1	1	79
22	chr2	104374659	107211278	0	0	1	27
23	chr2	111890936	118244338	0	1	1	70
24	chr2	120764979	120833500	0	1	0	1
25	chr2	126240290	127110286	0	1	1	6
26	chr2	133138750	149941557	1	1	1	215
27	chr2	157461751	157483350	0	1	0	1
28	chr2	167971120	168003596	0	1	0	2
29	chr2	185293453	185305296	1	0	0	1
30	chr2	229346615	229355363	1	0	0	1
31	chr3	33560667	35750525	1	0	0	39
32	chr3	54372654	54403213	1	0	0	0
33	chr3	98358330	98391049	0	1	0	0
34	chr3	118149134	118294301	1	0	0	3
35	chr4	18677194	18683044	1	0	0	1
36	chr4	21278849	21337830	0	1	0	2
37	chr4	26373871	26423688	0	1	0	3

38	chr4	31360987	34558254	0	1	1	52
39	chr4	45849344	47615468	0	1	1	12
40	chr4	59868197	60425643	0	1	1	8
41	chr4	70497566	74434332	1	0	1	46
42	chr4	82008572	82215823	0	1	0	6
43	chr4	90311130	91962113	0	0	1	8
44	chr4	121554377	124356978	0	1	1	27
45	chr4	128919570	133670654	1	0	0	48
46	chr4	136905695	136928673	0	0	1	0
47	chr4	143240267	147090538	1	0	0	59
48	chr4	184169869	184450981	1	0	0	7
49	chr4	203316217	203332954	1	0	0	0
50	chr4	224661121	224666412	1	0	0	0
51	chr5	1591098	1611277	0	1	0	3
52	chr5	11820020	11823805	1	0	0	1
53	chr5	67479937	67630316	1	0	0	5
54	chr5	87886831	89078200	1	0	0	18
55	chr5	92352752	92644615	0	1	0	2
56	chr5	125437902	126041703	0	0	1	4
57	chr5	147926211	147981089	0	1	0	1
58	chr5	160128174	164541434	1	1	1	81
59	chr5	192685362	192867950	1	0	1	3
60	chr5	201980075	201986492	1	0	0	1
61	chr6	39475298	39536174	1	0	0	0
62	chr6	70833345	75128389	1	0	0	67
63	chr6	104020474	104846933	1	0	0	14
64	chr6	111743312	111905380	1	0	0	3
65	chr6	118700694	119737122	1	0	0	16
66	chr6	132125931	132306220	1	0	0	7
67	chr6	136744931	137145831	0	0	1	8
68	chr6	138564837	138585701	1	0	0	1
69	chr6	144188428	144399049	0	1	0	2
70	chr6	148547648	150068199	1	0	0	39
71	chr6	160413038	165284918	1	0	0	233
72	chr7	9710307	13889417	0	1	1	87
73	chr7	17688939	17714548	1	0	0	2
74	chr7	44745316	46501661	0	1	1	25
75	chr7	146699976	148403614	1	0	0	48
76	chr7	165464112	165470747	1	0	0	2
77	chr8	30179302	30251700	1	0	0	1
78	chr8	37221664	42322860	1	0	0	57
79	chr8	61632813	64896557	1	0	0	55
80	chr8	71301175	71431155	1	0	0	1

81	chr8	90964543	94385148	1	0	0	51
82	chr8	105466415	105566934	1	0	0	2
83	chr8	113063388	114015951	1	0	0	19
84	chr8	119830324	119832288	1	0	0	1
85	chr8	133152921	134822866	1	0	0	46
86	chr8	142114111	142915956	1	0	0	19
87	chr8	170308877	171403851	1	0	0	49
88	chr9	384383	408227	1	0	0	0
89	chr9	20905213	21728170	0	1	1	14
90	chr9	24459413	30070094	1	0	0	130
91	chr9	110988581	110997371	1	0	0	1
92	chr9	120081323	120160910	1	0	0	2
93	chr9	147488930	147492242	0	1	0	1
94	chr9	149518807	149547799	1	0	0	6

Table S2 Candidate regions under selection during 30 generations of selection for seed size, at the 99.99% level. Regions were identified using F_{ST} values and a 25-single nucleotide polymorphism (SNP) sliding window approach. Comparisons were made between Krug Yellow Dent and KLS_30, Krug Yellow Dent and KSS_30, and KLS_30 and KSS_30. One indicates a difference and zero indicates no difference for the region.

Region	Chromosome	Start	End	Krug Yellow Dent vs. KLS_30	Krug Yellow Dent vs. KSS_30	KLS_30 vs. KSS_30	Number of Genes in Region
1	chr1	26329612	26830886	1	0	0	13
2	chr1	241368710	241403853	1	0	0	3
3	chr10	133216883	133233948	1	0	0	1
4	chr2	67171728	71897890	0	1	1	43
5	chr2	81659356	88321220	0	1	1	57
6	chr2	133888415	140323700	0	1	1	71
7	chr2	149509536	149793812	0	1	0	3
8	chr3	35626227	35655007	1	0	0	0
9	chr4	33053660	33128631	0	1	1	3
10	chr4	46050068	46061870	0	1	0	0
11	chr4	121594579	121609805	0	0	1	1
12	chr4	124305534	124320863	0	1	0	0
13	chr5	160954183	160971691	0	0	1	0
14	chr6	74962790	75080845	1	0	0	2
15	chr6	104456206	104843865	1	0	0	7
16	chr6	111761479	111767828	1	0	0	1
17	chr6	118702716	119665910	1	0	0	15
18	chr6	149827936	149835542	1	0	0	1
19	chr6	160589531	160606591	1	0	0	3
20	chr7	9901060	11800787	0	0	1	38
21	chr8	37229750	39230104	1	0	0	24
22	chr8	113178318	114007931	1	0	0	16
23	chr9	20905875	20973896	0	0	1	3

Table S3 Number of base pairs in the 2.1Gb maize v2 reference assembly with a given coverage range for each of the population pools. M=million.

	Population		
Coverage	Krug Yellow Dent	KLS_30	KSS_30
0	759M	859M	792M
1-5	523M	568M	447M
6-10	219M	225M	183M
11-15	144M	143M	127M
15-20	106M	99M	100M
21-25	81M	68M	82M
26-30	63M	45M	69M
31-40	86M	43M	105M
41-50	47M	12M	70M
>51	37M	4M	89M

Table S4 Genes within candidate regions under selection at the 99.9% level that were in a gene coexpression network module that distinguished KLS_30 and KSS_30 derived inbred lines and was enriched with cell cycle genes.

Chr	Start	End	99.9% Level Region	Krug Yellow Dent vs. KLS_30	Krug Yellow Dent vs. KSS_30	KLS_30 vs. KSS_30	Gene in Coexpression Module
chr1	198652274	203086089	7	1	0	0	GRMZM2G055968
chr1	260539402	263706680	11	1	0	0	GRMZM2G351304
chr2	79033475	88346653	21	0	1	1	GRMZM2G177596
chr2	104374659	107211278	22	0	0	1	GRMZM2G141814
chr2	133138750	149941557	26	1	1	1	GRMZM2G006765
chr2	133138750	149941557	26	1	1	1	GRMZM2G042897
chr4	70497566	74434332	41	1	0	1	GRMZM2G147756
chr4	128919570	133670654	45	1	0	0	GRMZM2G087323
chr6	118700694	119737122	65	1	0	0	GRMZM2G159953
chr6	160413038	165284918	71	1	0	0	GRMZM2G096389
chr6	160413038	165284918	71	1	0	0	GRMZM2G310758
chr6	160413038	165284918	71	1	0	0	GRMZM5G892879
chr7	9710307	13889417	72	0	1	1	GRMZM2G101036
chr7	9710307	13889417	72	0	1	1	GRMZM2G446921
chr7	146699976	148403614	75	1	0	0	AC196961.2_FG003
chr8	37221664	42322860	78	1	0	0	GRMZM2G120202
chr8	170308877	171403851	87	1	0	0	GRMZM2G069078
chr9	24459413	30070094	90	1	0	0	GRMZM2G050329
chr9	24459413	30070094	90	1	0	0	GRMZM2G136838

Table S5 Regions with copy number variation (CNV) between KLS_30 and KSS_30 based on read depth variation. Average read depth was determined in 5kb windows in both populations. CNV windows were defined as having an absolute value greater than 2 for the number of standard deviations (SD) away from the mean in KLS_30 minus the number of standard deviations away from the mean in KSS_30.

Chr	Region Start	Region Stop	Krug Yellow Dent SD From Mean	KLS_30 SD from Mean	KSS_30 SD From Mean	Absolute Value of KLS_30 SD Minus KSS_30 SD
1	235001	240000	6.20	3.54	5.81	2.27
1	203910001	203915000	32.13	36.10	29.50	6.60
1	234470001	234475000	9.85	6.52	8.83	2.31
1	234500001	234505000	8.74	5.46	8.02	2.56
1	234510001	234515000	10.74	6.57	9.38	2.81
1	234525001	234530000	6.38	3.88	6.34	2.46
1	234545001	234550000	2.00	7.27	0.28	6.99
1	234605001	234610000	19.13	13.87	16.02	2.15
1	234640001	234645000	11.68	6.30	10.07	3.76
1	234645001	234650000	23.55	10.47	20.49	10.02
1	234650001	234655000	4.62	1.99	4.03	2.04
1	234720001	234725000	30.96	17.98	26.04	8.05
1	234725001	234730000	32.95	19.19	27.93	8.74
1	234730001	234735000	26.88	15.52	22.49	6.97
1	234735001	234740000	22.14	13.38	18.65	5.27
2	65000001	65005000	9.96	11.93	9.33	2.60
2	77820001	77825000	9.52	5.09	8.19	3.10
2	77825001	77830000	20.88	9.50	18.85	9.35
2	77865001	77870000	8.97	5.59	7.65	2.06
2	77870001	77875000	24.74	16.35	22.08	5.73
2	77875001	77880000	30.42	18.85	26.10	7.25
2	77880001	77885000	13.41	7.95	11.53	3.57
2	172080001	172085000	31.16	35.19	31.35	3.84
2	172085001	172090000	14.78	9.31	14.01	4.70
2	172110001	172115000	26.79	30.90	25.12	5.78
2	172115001	172120000	53.85	60.10	51.33	8.77
2	174415001	174420000	1.82	5.69	-0.06	5.74
3	74660001	74665000	11.29	4.59	9.86	5.27
3	209600001	209605000	5.97	4.05	6.10	2.05
4	111670001	111675000	2.75	9.63	0.53	9.10
4	172415001	172420000	9.54	6.01	9.02	3.01
5	189240001	189245000	7.91	4.43	6.65	2.22
5	209940001	209945000	23.61	24.51	20.82	3.69
5	209945001	209950000	22.73	24.76	19.44	5.32
5	209960001	209965000	3.94	5.56	-0.32	5.88
5	209990001	209995000	19.42	22.00	18.48	3.52
5	210290001	210295000	4.68	6.30	4.11	2.18

6	20610001	20615000	7.53	6.25	8.27	2.02
6	60760001	60765000	9.35	11.79	8.53	3.26
6	104230001	104235000	17.50	12.94	16.65	3.71
6	160755001	160760000	10.34	7.23	9.34	2.11
6	160765001	160770000	68.65	72.54	63.99	8.55
6	160770001	160775000	33.42	27.46	30.52	3.06
6	160785001	160790000	12.43	5.81	11.82	6.01
7	18050001	18055000	9.32	11.19	8.66	2.53
7	44725001	44730000	25.12	29.51	22.88	6.63
8	80365001	80370000	8.94	5.13	8.89	3.75
8	97340001	97345000	1.68	1.18	3.21	2.03
8	97350001	97355000	2.11	1.61	3.85	2.23
8	146460001	146465000	9.07	10.99	8.45	2.54
9	6950001	6955000	21.67	14.18	17.76	3.58
9	6955001	6960000	17.92	11.16	15.19	4.03
9	57980001	57985000	11.86	5.65	10.62	4.97
9	67980001	67985000	14.12	9.01	12.19	3.18
9	68025001	68030000	19.43	13.21	17.14	3.93
10	34105001	34110000	27.75	13.75	22.70	8.95
10	121000001	121005000	16.96	10.89	15.30	4.41

Tables S6-S7

Available for download as Excel files at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.167155/-/DC1

Table S6 Comparative genome hybridization (CGH) normalized intensities for four inbreds generated from KLS_30 and five inbreds generated from KSS_30.

Table S7 Joint linkage analysis results for 20-kernel seed weight in the maize nested association mapping (NAM) population.

Table S8 Single nucleotide polymorphisms (SNPs) contained in a single forward regression genome wide association analysis (GWAS) model for 20-kernel seed weight in the maize nested association mapping (NAM) population. Effect is relative to B73.

Marker	chr	AGPv2 Position	cM	Effect	P value
PZE0123124739	1	23,099,268	39.91	-0.07	1.11E-16
PZE01201470169	1	201,639,860	115.12	-0.11	1.24E-09
PZE01237261221	1	237,965,327	140.05	-0.11	3.95E-15
PZE0207620470	2	7,663,333	22.98	-0.06	5.77E-17
PZE0219925121	2	20,005,607	50.23	0.15	1.15E-10
PZE0228682197	2	28,761,283	60.29	0.06	2.21E-17
PZE02207653607	2	210,665,344	116.70	0.11	2.08E-09
PZE0305630836	3	5,850,072	21.63	0.09	1.39E-08
PZE03182929802	3	184,677,342	94.84	0.13	1.35E-15
PZE03209569396	3	211,128,687	120.01	-0.16	5.34E-13
PZE04207608568	4	201,957,506	108.82	0.08	2.04E-20
PZE0545748962	5	46,424,011	59.89	-0.11	1.89E-15
PZE05209672404	5	210,474,175	129.99	0.25	1.87E-13
PZE0692901122	6	64,157,406	19.51	0.08	2.05E-09
PZE06159136863	6	159,014,181	80.65	-0.08	1.20E-10
PZE07148539524	7	154,191,204	86.44	0.09	1.15E-08
PZE07156647853	7	162,259,418	99.46	0.06	4.55E-14
PZE08103597003	8	104,822,548	58.93	-0.05	3.20E-13
PZE0961486830	9	NA	45.68	0.19	4.67E-24
PZE09131782056	9	136,179,349	66.59	0.06	2.23E-12
PZE1025657301	10	25,657,714	35.72	-0.09	7.26E-10

Table S9 Resampling model inclusion probability (RMIP) analysis results for 20-kernel seed weight in the maize nested association mapping (NAM) population. Only markers with bootstrap support in five or more subsamples are reported. Effect is relative to B73. The reported *P* values are the lowest significant *P* value that was observed across the 100 subsamples.

Marker	Chr	AGPv2 Position	cM	RMIP	Effect	P value
PZE0122275486	1	22,247,033	39.19	16	-0.12	6.15E-13
PZE0123077638	1	23,067,629	39.87	11	-0.07	8.38E-13
PZE0123124739	1	23,099,268	39.91	16	-0.07	5.61E-12
PZE0123662144	1	23,566,708	40.47	25	-0.07	5.02E-11
PZE0125025863	1	24,931,627	41.89	8	-0.08	6.15E-13
PZE0139180321	1	39,111,110	57.09	5	0.07	1.51E-09
PZE01201470169	1	201,639,860	115.12	9	-0.12	4.18E-09
PZE01233561761	1	234,219,193	138.59	39	-0.14	3.91E-11
PZE01237261221	1	237,965,327	140.05	40	-0.11	6.16E-11
PZE01292560885	1	293,627,855	192.50	12	-0.06	5.79E-10
PZE01292868532	1	293,935,502	193.21	5	-0.05	1.81E-09
PZE0205818953	2	5,817,525	17.74	8	-0.07	3.99E-11
PZE0207620470	2	7,663,333	22.98	48	-0.06	6.57E-11
PZE0207910201	2	7,953,064	23.73	9	-0.07	6.71E-11
PZE0219925121	2	20,005,607	50.23	18	0.16	1.84E-09
PZE0221648470	2	21,726,433	52.73	6	0.09	2.99E-09
PZE0228682191	2	28,761,277	60.29	6	0.07	3.63E-11
PZE0228682197	2	28,761,283	60.29	14	0.07	1.54E-13
PZE0229550868	2	29,117,510	61.21	6	0.06	2.48E-10
PZE0235758316	2	35,272,110	64.57	9	0.08	2.81E-14
PZE0238058171	2	37,572,981	66.09	6	0.07	5.49E-16
PZE0239176813	2	38,696,485	66.82	8	0.10	1.85E-10
PZE0240222660	2	39,757,715	67.51	29	0.11	5.75E-09
PZE0240904916	2	40,439,971	67.94	6	0.08	6.88E-11
PZE02207653607	2	210,665,344	116.70	6	0.13	1.99E-09
PZE0302919491	3	2,957,042	8.69	9	0.06	2.46E-09
PZE0305630836	3	5,850,072	21.63	8	0.10	8.31E-09
PZE03116146291	3	119,926,252	59.37	13	0.10	6.65E-09
PZE03177053561	3	178,806,797	88.27	6	0.15	1.15E-11
PZE03178447133	3	180,203,027	90.04	12	0.11	7.81E-12
PZE03182929802	3	184,677,342	94.84	65	0.13	2.99E-10
PZE03209569396	3	211,128,687	120.01	77	-0.16	1.39E-09
PZE04207608568	4	201,957,506	108.82	48	0.08	5.12E-13
PZE04207758758	4	202,107,696	108.86	23	0.07	3.74E-14
PZE04212652195	4	207,024,058	110.04	5	0.07	1.32E-14
PZE0536484165	5	37,174,222	58.23	14	-0.15	4.28E-12
PZE0545435902	5	46,110,951	59.83	10	-0.15	1.70E-10
PZE0545748962	5	46,424,011	59.89	21	-0.11	6.11E-12

PZE0566973506	5	67,673,484	64.68	6	-0.08	1.42E-11
PZE0567955527	5	68,647,181	64.90	6	-0.13	5.99E-13
PZE0570378999	5	71,092,575	65.73	9	-0.11	6.41E-12
PZE05209219847	5	210,021,618	128.36	9	0.18	8.04E-11
PZE05209416262	5	210,218,033	129.07	8	0.26	5.92E-11
PZE05209450970	5	210,252,741	129.19	5	0.26	1.84E-10
PZE05209890414	5	210,694,868	130.78	14	0.19	1.00E-10
PZE05212784052	5	213,583,963	142.19	20	0.19	8.37E-11
PZE05213906088	5	214,718,607	147.42	11	0.19	5.70E-10
PZE0690543233	6	91,646,020	17.51	7	0.08	4.93E-09
PZE0692901122	6	64,157,406	19.51	12	0.10	5.92E-09
PZE0696785554	6	96,541,043	22.94	6	0.15	1.08E-08
PZE06159136863	6	159,014,181	80.65	19	-0.08	6.37E-09
PZE06163919721	6	163,822,182	95.67	35	-0.09	5.30E-09
PZE07148539524	7	154,191,204	86.44	34	0.10	1.64E-08
PZE07156061393	7	161,697,119	98.24	6	0.09	1.04E-12
PZE07156647853	7	162,259,418	99.46	22	0.08	2.07E-10
PZE07157275574	7	162,985,624	100.76	18	0.09	6.83E-12
PZE07158131612	7	163,824,440	102.53	13	0.08	1.03E-11
PZE07160221189	7	165,945,883	107.21	13	0.06	5.46E-10
PZE07168993370	7	174,761,756	134.00	5	0.20	5.90E-09
PZE0801360932	8	1,375,719	2.97	9	-0.13	1.11E-08
PZE0832831580	8	32,859,653	51.21	5	-0.07	8.73E-12
PZE08103155726	8	104,380,896	58.78	12	-0.06	6.77E-11
PZE08103597003	8	104,822,548	58.93	12	-0.06	5.10E-11
PZE08109869427	8	111,192,862	60.88	9	-0.06	7.05E-12
PZE08112249901	8	113,634,215	61.49	7	-0.06	3.90E-11
PZE08156324673	8	157,638,136	83.03	6	-0.06	6.12E-11
PZE0961486830	9	NA	45.68	56	0.19	1.76E-15
PZE0985093978	9	88,002,312	46.49	5	0.22	9.81E-11
PZE0986885631	9	89,813,289	46.80	19	0.20	2.14E-13
PZE0988184281	9	91,122,357	47.01	6	0.20	3.83E-19
PZE09131781985	9	136,179,278	66.59	9	0.06	2.98E-09
PZE09131782056	9	136,179,349	66.59	9	0.06	3.98E-09
PZE09137421592	9	141,828,934	76.37	5	0.08	9.82E-09
PZE1025657301	10	25,657,714	35.72	25	-0.09	1.44E-08
PZE1030835021	10	30,870,720	36.00	17	-0.09	1.38E-08
PZE1036843968	10	52,676,288	36.08	6	-0.07	1.21E-08