the plant journal

SEB Society for Experimental Biology

The Plant Journal (2015) **84,** 587–596 doi: 10.1111/tpj.13029

Extreme-phenotype genome-wide association study (XP-GWAS): a method for identifying trait-associated variants by sequencing pools of individuals selected from a diversity panel

Jinliang Yang[1,†], Haiying Jiang[1,†,‡], Cheng-Ting Yeh[1], Jianming Yu[1], Jeffrey A. Jeddeloh[2], Dan Nettleton[3] and Patrick S. Schnable[1,4,*]

[1]*Department of Agronomy, Iowa State University, Ames, IA 50011, USA,*
[2]*Technology Innovation, Roche NimbleGen, Madison, WI 53719, USA,*
[3]*Department of Statistics, Iowa State University, Ames, IA 50011, USA, and*
[4]*Center for Plant Genomics, Iowa State University, Ames, IA 50011, USA*

Received 31 March 2015; revised 17 July 2015; accepted 8 September 2015; published online 12 September 2015.
*For correspondence (e-mail schnable@iastate.edu).
†These authors contributed equally to this work.
‡Present address: Shenyang Agricultural University, College of Agronomy, Shenyang 110161, China.

SUMMARY

Although approaches for performing genome-wide association studies (GWAS) are well developed, conventional GWAS requires high-density genotyping of large numbers of individuals from a diversity panel. Here we report a method for performing GWAS that does not require genotyping of large numbers of individuals. Instead XP-GWAS (extreme-phenotype GWAS) relies on genotyping pools of individuals from a diversity panel that have extreme phenotypes. This analysis measures allele frequencies in the extreme pools, enabling discovery of associations between genetic variants and traits of interest. This method was evaluated in maize (*Zea mays*) using the well-characterized kernel row number trait, which was selected to enable comparisons between the results of XP-GWAS and conventional GWAS. An exome-sequencing strategy was used to focus sequencing resources on genes and their flanking regions. A total of 0.94 million variants were identified and served as evaluation markers; comparisons among pools showed that 145 of these variants were statistically associated with the kernel row number phenotype. These trait-associated variants were significantly enriched in regions identified by conventional GWAS. XP-GWAS was able to resolve several linked QTL and detect trait-associated variants within a single gene under a QTL peak. XP-GWAS is expected to be particularly valuable for detecting genes or alleles responsible for quantitative variation in species for which extensive genotyping resources are not available, such as wild progenitors of crops, orphan crops, and other poorly characterized species such as those of ecological interest.

Keywords: extreme-phenotype genome-wide association study, exome-sequencing, trait-associated variants, diversity panel, maize, kernel row number.

INTRODUCTION

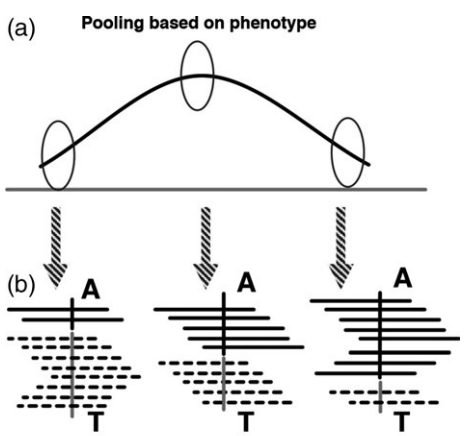
Despite the development of quantitative trait loci (QTL) mapping (Morton, 1955) and genome-wide association studies (GWAS) (Klein *et al.*, 2005), rapid and cost-effective identification of SNPs or genes associated with variation in complex traits remains challenging. Conventional QTL mapping is typically performed on the basis of newly occurring recombination events in the progeny of bi-parental crosses. Typically, at most only two to four alleles segregate in such crosses, limiting the number of trait-associated loci that may be detected. In addition, the limited number of recombination events usually results in relatively large confidence intervals. GWAS looks for associations using a greater fraction of the genetic diversity within a species that contributes to the trait of interest. When performed on large diversity panels, GWAS provides higher-resolution mapping of trait-associated variants (TAVs) because it exploits historical recombination events. However, one of the limitations of the

existing QTL mapping and GWAS approaches is that they require genotyping of large numbers of individuals, which may be expensive for large populations, even using recently developed cost-effective genotyping methods such as genotyping arrays (Steemers *et al.*, 2006; Fu *et al.*, 2010) and genotyping-by-sequencing (Elshire *et al.*, 2011).

An alternative method for identification of TAVs is bulk segregant analysis (BSA), which involves genotyping of pools of individuals sorted by phenotype rather than genotyping individuals within a segregating population or a diversity panel (Michelmore *et al.*, 1991). BSA may be performed using any type of genetic marker that provides a quantitative read-out that is correlated with allele frequencies in the phenotypically distinct pools. New implementations of BSA have recently been reported that exploit advances in genotyping technologies, especially the development of next-generation sequencing (NGS). For example, NGS-based BSA methods that rely on whole-genome shotgun sequencing have been applied to species with small genomes such as Arabidopsis (Schneeberger *et al.*, 2009) and *Saccharomyces cerevisiae* (Wenger *et al.*, 2010). Because these methods are not suitable for species with large genomes, we developed Sequenom-based BSA (Liu *et al.*, 2010) and RNA-seq based BSA (BSR-Seq) (Liu *et al.*, 2012); these technologies have been used by us and others to map or clone several maize genes whose qualitative mutants have large effects (Yi *et al.*, 2011; Makarevitch *et al.*, 2012; Li *et al.*, 2013). Similarly, sequencing-based strategies, such as next-generation mapping (Austin *et al.*, 2011), MutMap (Abe *et al.*, 2012) or mapping-by-sequencing (Mascher *et al.*, 2014) have been developed to detect point mutations in bi-parental populations. The extension of BSA to quantitative traits was demonstrated in a bi-parental cross of yeast (Ehrenreich *et al.*, 2010).

As is the case with QTL mapping studies, all these NGS-based BSA studies analyzed bi-parental populations that were segregating for only a fraction of the genetic diversity within a species. We wished to extend the NGS-based BSA approach to diversity panels to more fully sample and interrogate the genetic diversity that controls quantitative traits within a species. We were encouraged in this effort by the results of a simulation study that indicated that, if a sufficient number of progeny were used, NGS-based BSA was able to detect even small-effect loci (Ehrenreich *et al.*, 2010). In addition to reducing the number of samples to be genotyped, use of a pooling strategy has the potential to enrich for rare alleles and augment allele effects via extreme phenotypic selection. Hence, we elected to sequence pools of individuals that exhibit extreme phenotypes from a large diversity panel that contains historical recombination events. This method combines the simplicity of genotyping pools with the superior mapping resolution of GWAS (Figure 1), and was thus termed extreme-phenotype genome-wide association study (XP-GWAS).

We performed XP-GWAS for the quantitative trait kernel row number (KRN) of maize (*Zea mays*) using a diversity panel of approximately 7000 accessions. This trait was selected to enable comparisons of our results with those from a conventional GWAS (Brown *et al.*, 2011). Approximately 200 lines with the lowest KRNs and a similar number with the highest KRNs were selected from the diversity panel. In addition, a random set of approximately 200 lines from the diversity panel was used as a control. These three pools were genotyped via an exome capture and sequencing strategy that provided quantitative allele frequencies. XP-GWAS identified 145 TAVs. These variants are enriched in regions previously detected via traditional GWAS (Brown *et al.*, 2011). We also demonstrated the resolution of XP-GWAS by separating multiple linked QTL and identifying a single candidate gene under a single QTL peak. XP-GWAS combines BSA's simple experimental design with the high mapping resolution of GWAS, and may be particularly attractive for researchers studying species for which large, individually genotyped diversity panels do not exist or cannot easily be generated, such as orphan crops or ecological species.



(c)

Variant	Low	Random	High
Ref(A)	20	50	80
Alt(T)	80	50	20
Total	100	100	100

Figure 1. Simplified scheme for XP-GWAS by investigating the allele frequencies in different phenotypic pools.
(a) Based on the phenotypic distribution in a diversity panel, accessions with high, low or random phenotypes were pooled for sequencing.
(b) *De novo* variant discovery was performed, and the number of reads supported for reference variant call and alternative variant call were computed in each of the phenotypic pools.
(c) The variant counts at each locus were subjected to statistical testing.

RESULTS

Identification and pooling of lines with extreme KRN phenotypes

The North Central Regional Plant Introduction Station (NCRPIS), which is part of the US National Plant Germplasm System (NPGS), maintains more than 10 000 accessions of maize germplasm from across the world, representing the vast diversity of this species (Vigouroux *et al.*, 2008). Phenotypic data, including KRN counts, are available for 6952 of these accessions via the Germplasm Resources Information Network (GRIN) database (http://www.ars-grin.gov/). The KRN trait is approximately normally distributed within this diversity panel, with a mean of 13.4 (Figure 2). Using these KRN data, we established three pools of accessions. The mean and median phenotypic values for the three pools are 8.7/9 (low KRN pool), 13.5/13 (random KRN pool) and 19.7/19 (high KRN pool). Each pool consists of approximately 200 accessions (selection intensity approximately 3%) (Table S1). The random pool was created in addition to the high and low extreme phenotypic pools to reflect background population allele frequencies. The selected accessions originated from approximately 60 countries on six continents.

Validation of phenotypic scores via replicated field trials

To test the reproducibility of the phenotypic data downloaded from the GRIN database, replicated field trials (Experimental procedures) were performed using a subset of accessions selected because they represent extreme KRN phenotypes in the GRIN database. The correlation (r) between the GRIN data and our measurements in the low KRN pool was only 0.27 (n = 16 accessions, P value = 0.3) and that for the high KRN pool was only 0.45 (n = 29 accessions, P value < 0.01) (Figure S1). The reason for these low correlations is probably that individual accessions were both highly heterozygous and genetically heterogeneous. Even though the within-pool phenotypic correlations were relatively low, a high correlation was observed between the phenotypes of the two pools (r = 0.96, P value < 0.01). This indicates that members of the high KRN and low KRN pools may be clearly distinguished even using the phenotypes extracted from the GRIN database.

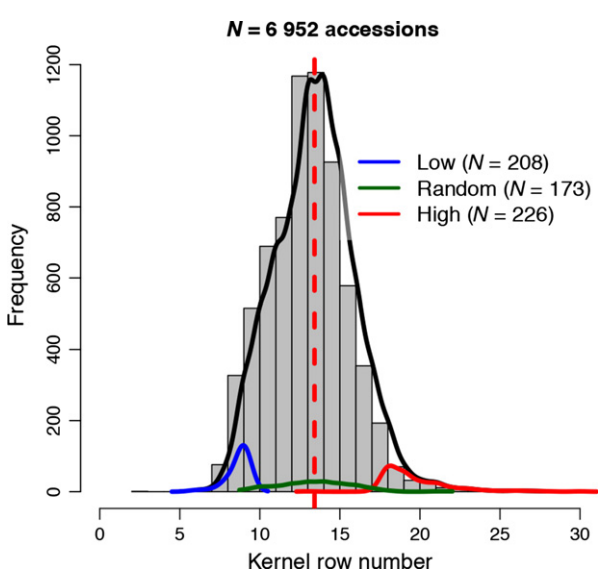


Figure 2. . KRN phenotype of diverse germplasm accessions.
Histogram and density plot of germplasm accessions (n = 6952) in the GRIN database, and density plots of three selected KRN phenotypic pools: low, high and random.

Exome-sequencing of three XP-GWAS pools

XP-GWAS starts with genotyping the extreme phenotype pools. Genotyping with a pre-defined SNP array creates an inherent ascertainment bias. This bias may be overcome by *de novo* SNP discovery within the pools, for example via whole-genome sequencing of each pool. However, because of its large genome (approximately 2.3 Gb) (Schnable *et al.*, 2009) and high proportion of repetitive DNA (approximately 80%) (Baucom *et al.*, 2009), we focused our sequencing resources on the genic regions of each pool. This was achieved by sequencing the products of an exome-capture experiment (Bashiardes *et al.*, 2005; Fu *et al.*, 2010)

A solution-based sequence capture library was designed and manufactured by NimbleGen (see Experimental procedures) to survey the complete B73 exome plus additional sequences that were not used in the current analysis. Using this 'Zeanome' probe library, sequence capture was performed on barcoded, fragmented genomic DNA samples from three XP-GWAS pools. The captured DNAs were then sequenced using four lanes of an Illumina HiSeq 2000 instrument, generating a total of approximately 770 million 100 bp paired-end reads. A custom bioinformatics pipeline (Li *et al.*, 2012) was used to align the raw reads to the maize B73 reference genome (RefGen_v2) (Experimental procedures). After data processing, approximately 302, 368 and 294 million single-end reads were uniquely mapped to the reference genome for the high, low and random KRN pools, respectively (Table S2). These uniquely mapped reads were analyzed to evaluate capture performance and to call variants.

The exome of the filtered gene set (FGSv5b.60) of the B73 reference genome was considered our intended target, although the design space included probes designed to other sequences (Experimental procedures). The mean sequencing depths on the filtered gene sets of the three pools were 142 x (high KRN), 175 × (low KRN) and 145 × (random KRN). Approximately 85% of the reference genes (84% for high KRN, 87% for low KRN and 84% for random KRN) have a depth of coverage greater than 50 × (Figure 3a–c). The mean percentages of coverage from

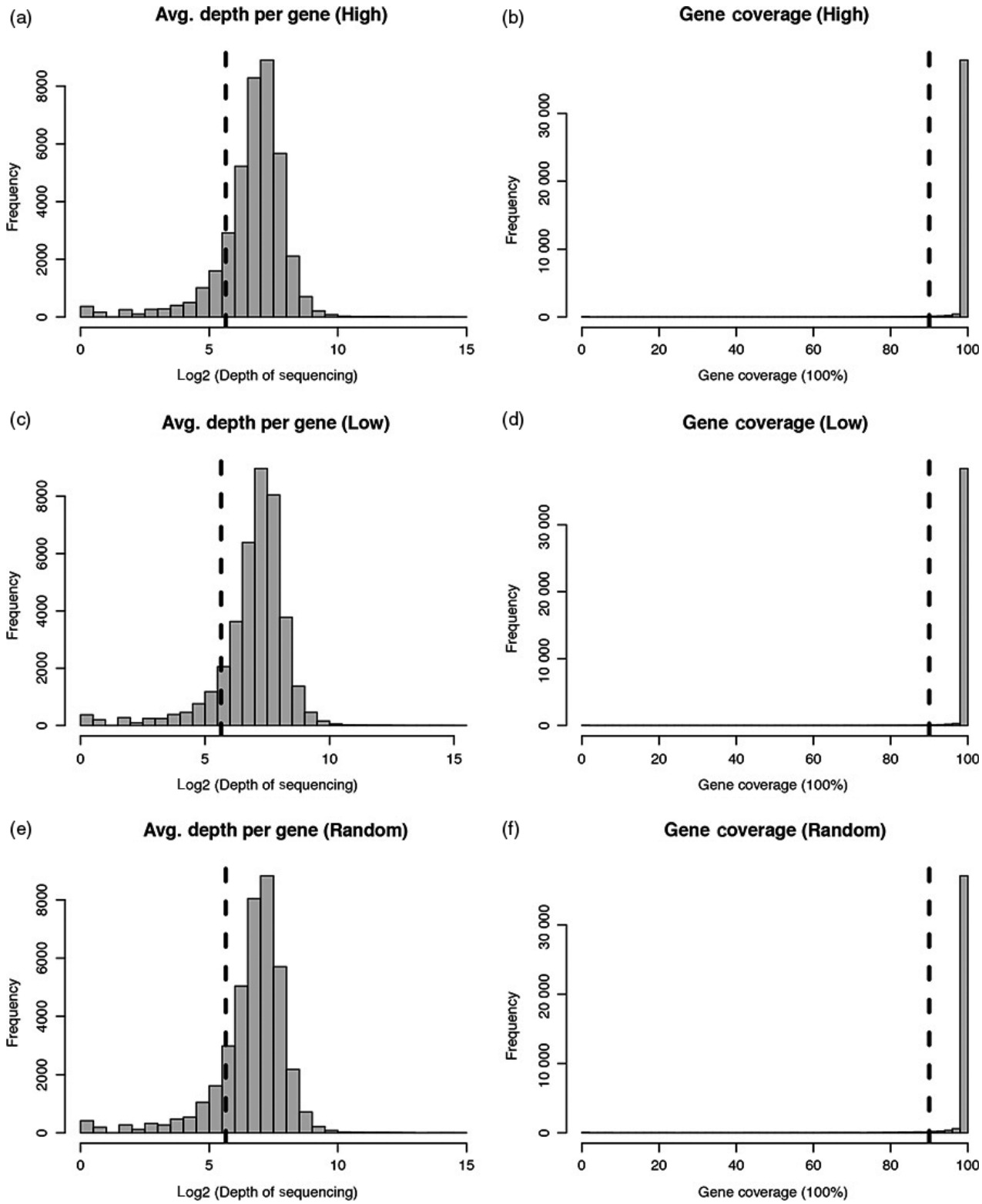


Figure 3. Histograms of depth of sequencing (a–c) and coverage (d–f) for the filtered gene set (FGSv5b.60) of three phenotypic pools.
(a–c) Vertical dashed lines indicate 50 x depth sequencing.
(d–f) Vertical dashed lines indicate 90% gene coverage.

transcript start to end for the reference genes were 99.0% (high KRN), 99.3% (low KRN) and 98.6% (random KRN). Approximately 98% of the reference genes (98% for high KRN, 99% for low KRN and 97% for random KRN) have at least 90% coverage (Figure 3d–f). Bait probes may capture adjacent regions (Fu *et al.*, 2010); therefore we anticipated capturing not only exonic regions but also intronic and promoter regions. Indeed, 7% of the reads (high KRN), 6% of the reads (low KRN) and 7% of the reads (random KRN) were mapped to intronic or 5 kb upstream regions. The results indicated that, even using conservative estimates, Zeanome Seq-Cap proved to be an efficient method to enrich the intended target with a high depth of sequencing and a high rate of coverage.

A total of 5.14 million variants including SNPs (n = 4.75 million, 92%) and small indels (n = 0.39 million, 8%) were identified using a custom variant calling pipeline (Experimental procedures). Simulation studies established that adequate read depth is critical to accurately estimate allele frequencies in the XP-GWAS pools (Figure S3). However, increasing the minimum read depth required to call variants dramatically reduces the number of common variants identified across the three pools (Figure S2). In addition, the read depth cut-off also affects the sensitivity for identification of rare alleles. Based on these simulations, we concluded that a minimum read depth of 50 × provides an appropriate balance between minimizing the negative effects of sampling variation and maintaining high numbers of variants detected. After filtering, using this 50 × read depth cut-off, 944 549 common variants were retained, including 828 855 SNPs (88%) and 115 694 small indels (12%). These variants were distributed across 87% of the high-confidence maize filtered genes (FGSv5b.60); on average, 18 variants were detected for each gene, and the most extreme gene (*GRMZM2G047347*) contains 246 polymorphic sites. As anticipated, variants were not limited to exonic regions; only approximately 41% of variants were located in exons. Approximately 34% of variants were located in introns, approximately 9% were located within 5 kb upstream of genes, and approximately 10% were located within 5 kb downstream of genes (Figure S4). This is relevant because, although genes and 5 kb upstream regions comprise only 13% of the genome, variations within these regions account for approximately 35–47% of phenotypic variation in maize (Li *et al.*, 2012). The ability of the Zeanome Seq-Cap library to capture both the exome and adjacent regions enabled us to focus sequencing resources, thereby enhancing the power of this study to identify associations.

Identification of extreme phenotype-associated variants

The primary factor used to create the three phenotyptic pools was the KRN phenotype. Even though an effort was made to maintain geographic diversity in the pools, population structure or cryptic within-group relatedness was unavoidable. This cryptic population structure may lead to over-dispersion of the χ^2 test statistic, thereby resulting in false discovery. To attenuate the effects of population structure, a genomic control method (Devlin and Roeder, 1999) was implemented to adjust the χ^2 test statistic (Experimental procedures). After implementing this genomic control, the quantile–quantile plot (Figure S5) showed that most of the observed data conformed closely to expectation except at the tail, indicating that the population structure was successfully controlled and some association signals were detected.

Using this approach, 145 TAVs were identified at a false discovery rate (FDR) (Benjamini and Hochberg, 1995) of 0.05 (Figure 4). These identified TAVs represent 121 1 kb bins distributed across ten chromosomes. To understand the patterns of differences in allele frequencies amongst the pools, at each TAV site read counts matching the reference allele were divided by total read counts to derive the reference allele frequency. We noted that the B73 inbred line (which provided the reference genome) with a mean of 17.6 kernel rows is phenotypically closer to the mean value of the high KRN pool (mean KRN = 19.5) than to the value of low KRN pool (mean KRN = 8.7), and previous studies found that the KRN trait was mostly controlled by additive gene effects (Toledo *et al.*, 2011). Therefore, it was not surprising that 81% (118/145) of reference allele frequencies of the TAVs exhibited an inheritance pattern of high > random > low, compared with only 1% (2/145) that exhibited the opposite pattern (high < random < low)

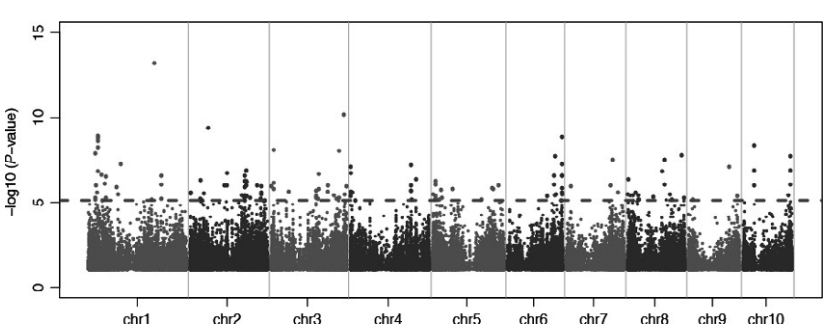


Figure 4. Manhattan plot of XP-GWAS results. The horizontal dashed line indicates the 5% FDR threshold.

(Figure S6a,c). The remaining TAVs exhibited other patterns (Figure S6b,d).

Comparisons between the results from XP-GWAS and traditional GWAS

We compared the 145 TAVs to 261 TAVs previously detected via conventional GWAS (Brown *et al.*, 2011). The two sets of variants were mapped to the same version of the reference genome (RefGen_v2). Using a bin size of 1 Mb, 17% of the TAVs (25/145) overlapped with the variants identified by traditional GWAS. This number of overlapping bins was statistically significant ($P < 0.05$) based on a simulation test (Experimental procedures). The TAVs were also compared with 986 recently identified TAVs (Yang *et al.*, unpublished results). Their study used 6230 entries (each of which was individually genotyped and phenotyped) from four related maize populations. The results of this traditional GWAS were analyzed using three complementary statistical approaches. Significant TAVs from these GWAS were subjected to cross-validation experiments in three independent populations. Using the same 1 Mb bins, 35% (51/145, P value < 0.05) of the TAVs identified in the present study overlapped with the TAVs identified via this 2nd conventional GWAS.

TAVs hit linked QTL regions with high resolution

Several previous QTL studies detected multiple QTLs that co-localize on the long arm of chromosome 4 (Beavis *et al.*, 1994; Veldboom *et al.*, 1994; Austin and Lee, 1996). A recent conventional study using GWAS (Yang *et al.*, unpublished results) also detected clusters of TAVs in these regions. In an attempt to resolve these linked QTLs, a chromosome walking experiment was performed (Experimental procedures), mapping one of the KRN QTL to a 2.7 Mb interval defined by a pair of SNPs (Figure S7).

XP-GWAS also identified TAVs in this region. Using pairwise comparisons of three independent χ^2 tests (high versus low, high versus random, and low versus random) with genomic control, four variants passed an FDR < 0.05 threshold and one variant was supported by two of the independent pairwise tests (high versus low and low versus random) (Figure 5a). These four TAVs were all located in the gene *GRMZM2G039106*, which is itself located under the peak of the fine-mapped QTL interval (Figures 5b and S7). The high KRN pool of these variants maintained high reference allele frequencies, consistent with our original determination that the favorable allele was derived from B73. In addition to identifying TAVs in this region, XP-GWAS identified TAVs in three other chromosomal

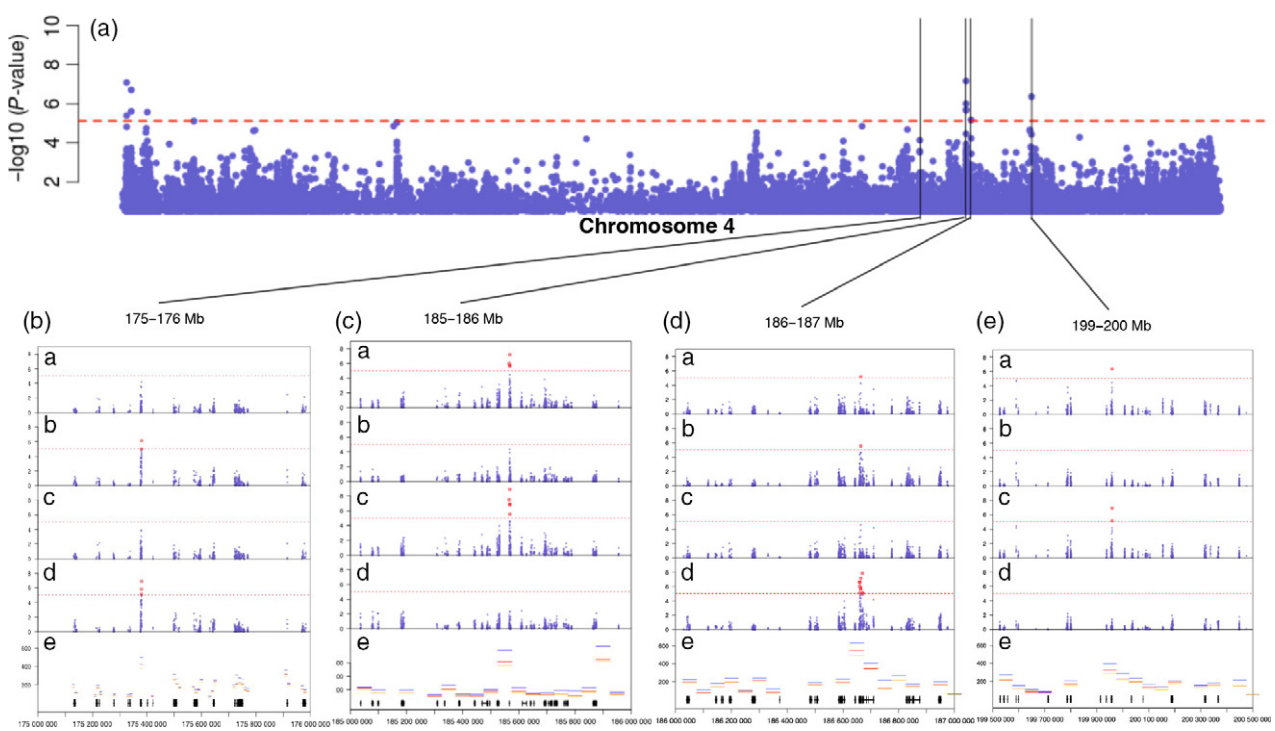


Figure 5. XP-GWAS and independent pairwise χ^2 test results on regions of chromosome 4.
(a) XP-GWAS results.
(b–e) Magnified results for four chromosomal regions. The top four panels show a magnified plot of the XP-GWAS in the region, and independent pairwise χ^2 tests for the high KRN pool versus the low KRN pool, the high KRN pool versus the random KRN pool, and the low KRN pool versus the random KRN pool, respectively. Red dashed lines indicate the 5% FDR threshold. The bottom panels show the read depth of three KRN pools using a bin size of 1000, where blue lines represent the high KRN pool, red lines represent the low KRN pool, and yellow lines represent the random KRN pool, respectively.

regions on the long arm of chromosome 4: chr4:185-186M, chr4:186-187M and chr4:200-201M (Figure 5c–e). These variants were located in genes *GRMZM2G111928*, *GRMZM2G095141* and *GRMZM2G098557*. Favorable alleles of these loci were all derived from B73, which is consistent with previous QTL findings.

DISCUSSION

Conventional GWAS experiments have been used to identify loci associated with important traits in agricultural species. These analyses require that large panels of individuals be genotyped, which is still expensive despite recent advances in genotyping technologies. Recently, regions of the maize genome under selection for a quantitative trait (seed size) were detected by sequencing pools of individuals from pairs of extreme populations derived from a long-term divergent selection program (Hirsch *et al.*, 2014). Although this approach eliminates the need to genotype large panels of individuals, it requires access to populations that have undergone multiple generations of selection. In contrast, XP-GWAS relies on the pooling of extreme phenotypes from readily available diversity panels, and may be applied to any trait of interest. Using XP-GWAS, we identified 145 TAVs with only several days of hands-on time and at modest cost. In this study, the resolution afforded by XP-GWAS was comparable with results from conventional GWAS. Specifically, using both methods, we were able to identify variants within a fine-mapped QTL region that is embedded within a cluster of linked QTL. As a consequence of the exome sequencing strategy used in this implementation of XP-GWAS, approximately 90% of the TAVs were located in genes and their 5 kb flanking regions. Although these identified genes are potential candidates for further investigations, as is the case for associations obtained from all GWAS, the identified TAVs and the associated genes may not themselves be causative but may have been identified as a consequence of linkage disequilibrium (LD).

The power of XP-GWAS is affected by many factors, including the precision of phenotyping, pool sizes, selection intensity, marker density, and the depth of sequencing. Our results demonstrate that XP-GWAS tolerates a degree of inaccuracy in the phenotyping data. For example, the KRN phenotypic data used in this study were collected based on observations during routine seed propagation activities at the North Central Regional Plant Introduction Station rather than via a systematic field trial design. However, XP-GWAS would be expected to have more power if the underlying phenotypic data were more precisely assayed. In addition, simulated power analyses for BSA found that increasing the bulk size while maintaining the selection intensity constant at 5% has the potential to increase the power to detect small-effect QTLs (Ehrenreich *et al.*, 2010). This implies that large diversity panels are desirable. Another simulation study (Magwene *et al.*, 2011) suggested that BSA would be more powerful if the selection intensity were higher than 10%, assuming sufficient amounts of quantitative genotyping. This report recommended using a mean depth of sequencing at least as high as the number of individuals in a pool. In addition to depth of sequencing, adequate maker density is also critical. The ability of XP-GWAS to detect associations relies on differences in allele frequencies of markers in LD with the QTL of interest in the phenotypic pools. Therefore, the appropriate marker density for XP-GWAS, just as for conventional GWAS, is affected by the size of LD blocks. Although LD is affected by many factors (Flint-Garcia *et al.*, 2009), LD generally decays rapidly in out-crossing plant species such as maize (range 1–10 kb) (Yan *et al.*, 2009), and more slowly in self-pollinating species such as rice (*japonica* approximately 150 kb; *indica* approximately 75 kb) (Mather *et al.*, 2007) and Arabidopsis (250 kb) (Nordborg *et al.*, 2002). In the 2.3 Gb maize genome, between 200 000 and 2 000 000 markers (2.3 Gb/1–10 kb) are required to capture most of the genomic variation, assuming an LD block size of 1–10 kb. The number of markers in the current study (944 549) is within this range. Estimates of the number of markers required for other species may be computed similarly.

The above considerations not only determine the power of the study, they may also inform decisions about the appropriate genotyping technologies to be used for XP-GWAS based on the size of the target species' genome and available resources. For relative small genomes such as cucumber (243.5 Mb) (Huang *et al.*, 2009) and strawberry (approximately 240 Mb) (Shulaev *et al.*, 2011), whole-genome sequencing may detect not only SNPs and short indels, but also present/absent variations (PAVs) and copy number variations (CNVs). For larger, complex genomes, options for reduced representation genotyping include restriction digestion-based methods (Van Tassell *et al.*, 2008), RNA sequencing (Haseneyer *et al.*, 2011) and targeted sequence capture (Bashiardes *et al.*, 2005). Some of these methods may be applied to species that lack reference genomes. If an RNA-seq based genotyping approach (Barbazuk *et al.*, 2007) is used, loci that exhibit associations to traits must be interpreted within the context of their expression profiles.

This study introduces the Zeanome capture library to the maize genetics toolkit. Using this library, it is possible to focus sequencing resources on the non-repetitive portions of this large and repetitive genome.

Although XP-GWAS has significant advantages compared to other methods of identifying marker/trait associations, it also has some inherent limitations. For example, as a consequence of the need to pool individuals by phenotypes, a separate XP-GWAS experiment must be performed for each trait of interest. Furthermore, because inferences rely on allele frequencies in populations, it is

probably not possible to estimate individual variant effects and heritability via XP-GWAS. The number of marker/trait associations detected by XP-GWAS ($n = 145$) in this experiment was somewhat lower than the number obtained via conventional GWAS ($n = 260$) (Brown *et al.*, 2011). However, because the two study populations have different genetic compositions, the absence of complete overlap between the two experiments may be at least partly a consequence of population-specific signals. In addition, even if the power of XP-GWAS is increased by using greater depth of sequencing, larger and/or better-designed pools and more precise phenotyping, XP-GWAS is not expected to yield better performance than conventional GWAS because pooling introduces stochastic effects and uncertainties. This limitation is counter-balanced by the substantial reduction in genotyping cost for XP-GWAS compared to conventional GWAS.

Conventional GWAS has the potential to yield false-positive signals as a consequence of population structure; this remains an important issue for XP-GWAS. False associations arise if a set of closely related lines are included in one extreme pool and another set of related lines are present in the other extreme pool. To reduce the effects of population structure, we introduced a random pool. Because this pool is a random sample of the population (i.e. the diversity panel), variant frequencies in this pool were treated as estimates of these frequencies in the population. Second, a statistical approach widely used in conventional GWAS was adapted to correct the inflation of the test statistic in XP-GWAS. In this method, a genomic control parameter λ was defined as the median (or mean) χ^2 association statistic across genome-wide markers divided by its theoretical value under the null distribution. A value of $\lambda = 1$ indicates the absence of population structure effects, while $\lambda > 1$ indicates the existence of some degree of population structure that should be corrected for.

In addition to the above approaches, a careful experimental design has the potential to reduce the effects of population structure. Similarly, use of matched samples (selecting pairs of extreme samples from the same geographic origin) has potential to reduce the effects of population structure. Of course this will not be effective if accessions from the same geographic region do not have similar genetic backgrounds, for example as a consequence of migration. To overcome this problem, ancestry matching through genotyping was proposed in a human case and control study (Crossett *et al.*, 2010). Type I error may be effectively controlled by this method, but requires genotyping of individual samples within the pools, albeit perhaps with only a small number of markers.

By taking advantage of advances in sequencing technologies and the development of appropriate statistical approaches, XP-GWAS promises to enhance the rate of genetic gain in crops, e.g. by identifying loci that may be used for marker-assisted selection and allele mining. XP-GWAS may also be used for the discovery of loci that play important roles in ecologically significant wild species, e.g. genes that confer resistance to stresses associated with climate change. Our initial XP-GWAS was performed in maize because it was possible to compare our results with those obtained from conventional GWAS. However, the most appropriate targets for XP-GWAS are probably not major crops such as maize, for which extensive previously genotyped diversity panels exist. Instead, XP-GWAS may be most relevant for minor/orphan crops (Collard and Mackill, 2008; Varshney *et al.*, 2012), for which large, phenotypically characterized germplasm collections often already exist. These existing phenotypic data may be used for XP-GWAS as an efficient and cost-effective method to identify loci that control agronomically significant loci.

EXPERIMENTAL PROCEDURES

Genetic materials, DNA extraction and phenotyping

Maize germplasm accessions were obtained from the North Central Regional Plant Introduction Station (NCRPIS, http://www.ars.usda.gov/main/) based on the KRN records in the GRIN database (http://www.ars-grin.gov/). Efforts were made to select geographically diverse lines. However, the majority of the accessions selected were from the USA. As a further control, we selected similar numbers of high, low and random accessions from the USA. A similar strategy was used for other countries for which sufficient numbers of accessions with sufficient phenotypic diversity were available. These accessions were bulked into three phenotypic pools: the high KRN pool ($n = 226$), low KRN pool ($n = 208$) and random KRN pool ($n = 173$) (Table S1). Because these accessions are both heterogeneous and heterozygous, tissue for DNA extraction was sampled from 12 plants per accession and pooled. After pooling tissues from all accessions, DNA was extracted using a CTAB method (Clarke, 2009).

To test the accuracy of the phenotypic data, 45 accessions (29 from the high KRN pool and 16 from the low KRN pool) were planted in two locations. Within each replication, each accession was planted in a row of 12 plants. KRN phenotypes were determined at harvest. Values of KRN were estimated by fitting a mixed linear model (Gilmour *et al.*, 1997), where genotype was fitted as a fixed effect and location was fitted as a random effect. The mixed model was implemented using the R add-on package 'nlme' (https://cran.r-project.org/web/packages/nlme/index.html).

Zeanome capture probe design and exome sequencing

A solution-based Zeanome capture library was designed by Roche Nimblegen. This library contains 186 513 probes designed from 39 656 maize gene models (FGSv5b.60) comprising approximately 60 Mb of non-repetitive sequences.

Before exome capture, indexed adapters (barcodes) were separately added to the three pooled DNA samples according to the TruSeq DNA sample preparation guide (Illumina). After DNA quantification using an Agilent bioanalyzer and a high-sensitivity DNA bioanalyzer kit (catalog number 5067–4626), 300 μg DNA from each sample were pooled. Then sequence capture was performed according to the NimbleGen protocol. The captured DNA was quantified on the Agilent bioanalyzer and diluted to

10 ng μl^{-1}. The prepared library was sequenced on an Illumina HiSeq 2000 machine using the paired-end 100-cycle protocol. The exome-capture sequence data have been deposited at the National Center for Biotechnology Information Short Read Archive under accession number SRP060571.

Genome alignment and variant calling

As reported in our previous study (Li *et al.*, 2012), the nucleotides of raw reads were scanned for low quality, and quality values lower than the threshold were trimmed using a custom pipeline. Trimmed reads were then aligned to the reference genome using GSNAP (Wu and Nacu, 2010) as paired-end fragments. The coordinates of confident and single (unique) alignment that passed our filtering criteria were used for SNP and small indel discovery. Polymorphisms at each potential variant site were carefully examined and putative variants were identified.

XP-GWAS with genomic control

To perform XP-GWAS, a generalized linear model analysis was performed for each variant using the 'glm' function in R. Each reference allele count for a given phenotypic pool was modeled as a binomial random variable with the number of trials equal to the sum of the reference and alternative allele counts for that pool. The logit of the binomial success probability was modeled as a linear function of phenotypic pool number (1 = low KRN, 2 = random KRN, 3 = high KRN). The likelihood ratio test statistic for testing the null model of no association between success probability and phenotypic pool number was computed. This statistic will tend to be large when the success probabilities either increase or decrease with phenotypic pool number. Due to the non-independence of samples raised from population structure, the usual χ^2 approximation to the distribution of each likelihood ratio test statistic is inappropriate. To correct for this, the inflation factor λ proposed by Devlin and Roeder (1999) was estimated using the R add-on package 'gap' (https://cran.r-project.org/web/packages/gap/index.html). The likelihood ratio test statistics were divided by λ and used to derive *P* values. Finally, the FDR method was applied to the resulting *P* values to identify significant variants while controlling FDR at the 5% level. Scripts associated with this study are available on Github (https://github.com/schnablelab/XP-GWAS).

Fine mapping of a QTL located in the Chr4:169-180 Mb interval

In an earlier study, we identified KRN QTL using 291 inter-mated B73 and Mo17 recombinant inbred lines (Yang *et al.* unpublished results). One large-effect QTL (effect = 1.3 rows, heritability = 15%) located in the Chr4:169-179 Mb interval was selected for fine mapping. Two SNP markers (M13783 and M89103) were designed to define the QTL interval. After genotyping a set of IBM recombinant inbred lines, two recombinant inbred lines (M0024 and M0054), which contain an Mo17 fragment on the QTL interval, were identified. These two recombinant inbred lines were back-crossed to B73 twice to create the F1BC2 mapping population, with the first back-cross being performed after genotyping. On average, the mapping population contained approximately 6% Mo17 materials in a B73 background. After screening approximately 6100 of these F1BC2 plants using the two SNP markers, 262 recombinants were initially identified, and they both selfed and outcrossed to their recurrent parent B73. These identified recombinants were further genotyped using 26 SNPs located within the QTL interval to define their recombination break points (Table S3). For the back-crossed recombinants, 26 heterozygous families with unique break points were chosen and planted in replicated plot trails with 12 plants in six replications to collect the KRN phenotype. For the selfed recombinants, 220 homozygous recombinant families were successfully created by further selfing the identified homozygous plants. These homozygous recombinants were phenotyped using a replicated field design with seven plants in eight replications (Table S4).

Monte Carlo simulation procedures

A Monte Carlo simulation procedure (Rubinstein, 1981) was used to test the hypothesis that the number of overlapping loci between this study and previous results via traditional GWAS has no difference than expected by chance. First, the same number of variants was sampled from a set of 0.94 million variants to resemble the TAVs. Then the number of overlap TAVs was recorded as the test statistic. This procedure was repeated 10 000 times, and the number of test statistics larger than the observation value was divided by the total number of simulations to derive an empirical Monte Carlo *P* value (Johnson *et al.*, 2011).

ACKNOWLEDGMENTS

We gratefully acknowledge Wei Wu, Heng-Cheng Hu and Talissa Sari for technical assistance, Lisa Coffey for germplasm management, and the US Department of Agriculture/Agricultural Research Service North Central Region Plant Introduction Station (NCRPIS) for providing germplasm. This research was supported by a grant from the US National Science Foundation to P.S.S. and colleagues (IOS-1027527). The Roche NimbleGen research and development group is privately funded.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Validation of KRN phenotypic values in the database using our replicated field trial observations.

Figure S2. Plots of number of variants and the required read depth for variant calling.

Figure S3. Simulation of random sampling error with different read depth.

Figure S4. Distribution of off-target variants identified by exome sequencing.

Figure S5. Quantile–quantile plots of the χ^2 distribution of the XP-GWAS.

Figure S6. Reference allele frequencies of the identified TAVs.

Figure S7. QTL fine mapping results.

Table S1. Accession IDs and KRN values of the selected high KRN, low KRN and random KRN lines.

Table S2. Summary of exome sequencing results of the three KRN pools.

Table S3. Genotypic data of the identified recombinants for QTL fine mapping of Chr4:169-180 Mb using 26 SNPs.

Table S4. KRN phenotypic values of the identified homozygous recombinant families.

REFERENCES

Abe, A., Kosugi, S., Yoshida, K. *et al.* (2012) Genome sequencing reveals agronomically important loci in rice using MutMap. *Nat. Biotechnol.* **30**, 174–178.

Austin, D.F. and Lee, M. (1996) Comparative mapping in F-2:3 and F-6:7 generations of quantitative trait loci for grain yield and yield components in maize. *Theor. Appl. Genet.* **92**, 817–826.

Austin, R.S., Vidaurre, D., Stamatiou, G. *et al.* (2011) Next-generation mapping of Arabidopsis genes. *Plant J.* **67**, 715–725.

Barbazuk, W.B., Emrich, S.J., Chen, H.D., Li, L. and Schnable, P.S. (2007) SNP discovery via 454 transcriptome sequencing. *Plant J.* **51**, 910–918.

Bashiardes, S., Veile, R., Helms, C., Mardis, E.R., Bowcock, A.M. and Lovett, M. (2005) Direct genomic selection. *Nat. Methods* **2**, 63–69.

Baucom, R.S., Estill, J.C., Chaparro, C., Upshaw, N., Jogi, A., Deragon, J.M., Westerman, R.P., Sanmiguel, P.J. and Bennetzen, J.L. (2009) Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome. *PLoS Genet.* **5**, e1000732.

Beavis, W.D., Smith, O.S., Grant, D. and Fincher, R. (1994) Identification of quantitative trait loci using a small sample of topcrossed and F4 progeny from maize. *Crop Sci.* **34**, 882–896.

Beissinger, T.M., Hirsch, C.N., Vaillancourt, B., Deshpande, S., Barry, K., Buell, C.R., Kaeppler, S.M., Gianola, D. and de Leon, N. (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.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300.

Brown, P.J., Upadyayula, N., Mahone, G.S. *et al.* (2011) Distinct genetic architectures for male and female inflorescence traits of maize. *PLoS Genet.* **7**, e1002383.

Clarke, J.D. (2009) Cetyltrimethyl ammonium bromide (CTAB) DNA miniprep for plant DNA isolation. *Cold Spring Harb. Protoc.* **2009**, pdb.prot5177. doi: 10.1101/pdb.prot5177.

Collard, B.C. and Mackill, D.J. (2008) Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **363**, 557–572.

Crossett, A., Kent, B.P., Klei, L., Ringquist, S., Trucco, M., Roeder, K. and Devlin, B. (2010) Using ancestry matching to combine family-based and unrelated samples for genome-wide association studies. *Stat. Med.* **29**, 2932–2945.

Devlin, B. and Roeder, K. (1999) Genomic control for association studies. *Biometrics*, **55**, 997–1004.

Ehrenreich, I.M., Torabi, N., Jia, Y., Kent, J., Martis, S., Shapiro, J.A., Gresham, D., Caudy, A.A. and Kruglyak, L. (2010) Dissection of genetically complex traits with extremely large pools of yeast segregants. *Nature*, **464**, 1039–1042.

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S. and Mitchell, S.E. (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE*, **6**, e19379.

Flint-Garcia, S.A., Buckler, E.S., Tiffin, P., Ersoz, E. and Springer, N.M. (2009) Heterosis is prevalent for multiple traits in diverse maize germplasm. *PLoS ONE*, **4**, e7433.

Fu, Y., Springer, N.M., Gerhardt, D.J. *et al.* (2010) Repeat subtraction-mediated sequence capture from a complex genome. *Plant J.* **62**, 898–909.

Gilmour, A.R., Cullis, B.R. and Verbyla, A.P. (1997) Accounting for natural and extraneous variation in the analysis of field experiments. *J. Agric. Biol. Environ. Stat.* **2**, 269–293.

Haseneyer, G., Schmutzer, T., Seidel, M. *et al.* (2011) From RNA-seq to large-scale genotyping – genomics resources for rye (*Secale cereale* L.). *BMC Plant Biol.* **11**, 131.

Hirsch, C.N., Flint-Garcia, S.A., Beissinger, T.M. *et al.* (2014) Insights into the effects of long-term artificial selection on seed size in maize. *Genetics*, **198**, 409–421.

Huang, S., Li, R., Zhang, Z. *et al.* (2009) The genome of the cucumber, *Cucumis sativus* L. *Nat. Genet.* **41**, 1275–1281.

Johnson, C., Drgon, T., Walther, D. and Uhl, G.R. (2011) Genomic regions identified by overlapping clusters of nominally-positive SNPs from genome-wide studies of alcohol and illegal substance dependence. *PLoS ONE* **6**, e19210.

Klein, R.J., Zeiss, C., Chew, E.Y. *et al.* (2005) Complement factor H polymorphism in age-related macular degeneration. *Science*, **308**, 385–389.

Li, X., Zhu, C., Yeh, C.T. *et al.* (2012) Genic and nongenic contributions to natural variation of quantitative traits in maize. *Genome Res.* **22**, 2436–2444.

Li, L., Li, D., Liu, S. *et al.* (2013) The maize *glossy13* gene, cloned via BSR-Seq and Seq-walking, encodes a putative ABC transporter required for the normal accumulation of epicuticular waxes. *PLoS ONE*, **8**, e82333. Correction: *PLoS ONE* **9**, e99563.

Liu, S., Chen, H.D., Makarevitch, I., Shirmer, R., Emrich, S.J., Dietrich, C.R., Barbazuk, W.B., Springer, N.M. and Schnable, P.S. (2010) High-throughput genetic mapping of mutants via quantitative single nucleotide polymorphism typing. *Genetics*, **184**, 19–26.

Liu, S., Yeh, C.T., Tang, H.M., Nettleton, D. and Schnable, P.S. (2012) Gene mapping via bulked segregant RNA-Seq (BSR-Seq). *PLoS ONE*, **7**, e36406.

Magwene, P.M., Willis, J.H. and Kelly, J.K. (2011) The statistics of bulk segregant analysis using next generation sequencing. *PLoS Comput. Biol.* **7**, e1002255.

Makarevitch, I., Thompson, A., Muehlbauer, G.J. and Springer, N.M. (2012) Brd1 gene in maize encodes a brassinosteroid C-6 oxidase. *PLoS ONE*, **7**, e30798.

Mascher, M., Jost, M., Kuon, J.E., Himmelbach, A., Assfalg, A., Beier, S., Scholz, U., Graner, A. and Stein, N. (2014) Mapping-by-sequencing accelerates forward genetics in barley. *Genome Biol.* **15**, R78.

Mather, K.A., Caicedo, A.L., Polato, N.R., Olsen, K.M., McCouch, S. and Purugganan, M.D. (2007) The extent of linkage disequilibrium in rice (*Oryza sativa* L.). *Genetics*, **177**, 2223–2232.

Michelmore, R.W., Paran, I. and Kesseli, R.V. (1991) Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. *Proc. Natl Acad. Sci. USA* **88**, 9828–9832.

Morton, N.E. (1955) Sequential tests for the detection of linkage. *Am. J. Hum. Genet.* **7**, 277–318.

Nordborg, M., Borevitz, J.O., Bergelson, J. *et al.* (2002) The extent of linkage disequilibrium in *Arabidopsis thaliana*. *Nat. Genet.* **30**, 190–193.

Rubinstein, R.Y. (1981) *Simulation and the Monte Carlo Method*. New York: Wiley.

Schnable, P.S., Ware, D., Fulton, R.S. *et al.* (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science*, **326**, 1112–1115.

Schneeberger, K., Ossowski, S., Lanz, C., Juul, T., Petersen, A.H., Nielsen, K.L., Jorgensen, J.E., Weigel, D. and Andersen, S.U. (2009) SHOREmap: simultaneous mapping and mutation identification by deep sequencing. *Nat. Methods* **6**, 550–551.

Shulaev, V., Sargent, D.J., Crowhurst, R.N. *et al.* (2011) The genome of woodland strawberry (*Fragaria vesca*). *Nat. Genet.* **43**, 109–116.

Steemers, F.J., Chang, W., Lee, G., Barker, D.L., Shen, R. and Gunderson, K.L. (2006) Whole-genome genotyping with the single-base extension assay. *Nat. Methods* **3**, 31–33.

Toledo, F.H.R.B., Ramalho, M.A.P., Abreu, G.B. and de Souza, J.C. (2011) Inheritance of kernel row number, a multicategorical threshold trait of maize ears. *Genet. Mol. Res.* **10**, 2133–2139.

Van Tassell, C.P., Smith, T.P., Matukumalli, L.K., Taylor, J.F., Schnabel, R.D., Lawley, C.T., Haudenschild, C.D., Moore, S.S., Warren, W.C. and Sonstegard, T.S. (2008) SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nat. Methods* **5**, 247–252.

Varshney, R.K., Ribaut, J.M., Buckler, E.S., Tuberosa, R., Rafalski, J.A. and Langridge, P. (2012) Can genomics boost productivity of orphan crops? *Nat. Biotechnol.* **30**, 1172–1176.

Veldboom, L.R., Lee, M. and Woodman, W.L. (1994) Molecular marker-facilitated studies in an elite maize population. 1. Linkage analysis and determination of QTL for morphological traits. *Theor. Appl. Genet.* **88**, 7–16.

Vigouroux, Y., Glaubitz, J.C., Matsuoka, Y., Goodman, M.M., Sanchez G, J. and Doebley, J. (2008) Population structure and genetic diversity of New World maize races assessed by DNA microsatellites. *Am. J. Bot.* **95**, 1240–1253.

Wenger, J.W., Schwartz, K. and Sherlock, G. (2010) Bulk segregant analysis by high-throughput sequencing reveals a novel xylose utilization gene from *Saccharomyces cerevisiae*. *PLoS Genet.* **6**, e1000942.

Wu, T.D. and Nacu, S. (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, **26**, 873–881.

Yan, J., Shah, T., Warburton, M.L., Buckler, E.S., McMullen, M.D. and Crouch, J. (2009) Genetic characterization and linkage disequilibrium estimation of a global maize collection using SNP markers. *PLoS ONE*, **4**, e8451.

Yi, G., Lauter, A.M., Scott, M.P. and Becraft, P.W. (2011) The *thick aleurone1* mutant defines a negative regulation of maize aleurone cell fate that functions downstream of *defective kernel1*. *Plant Physiol.* **156**, 1826–1836.