the plant journal

SEB Society for Experimental Biology

The Plant Journal (2014) **79,** 361–374 doi: 10.1111/tpj.12569

FEATURED ARTICLE

Genome diversity in *Brachypodium distachyon:* deep sequencing of highly diverse inbred lines

Sean P. Gordon[1], Henry Priest[2], David L. Des Marais[3], Wendy Schackwitz[4], Melania Figueroa[5,6], Joel Martin[4], Jennifer N. Bragg[1,7], Ludmila Tyler[8], Cheng-Ruei Lee[9], Doug Bryant[2], Wenqin Wang[10], Joachim Messing[10], Antonio J. Manzaneda[11], Kerrie Barry[4], David F. Garvin[12], Hikmet Budak[13], Metin Tuna[14], Thomas Mitchell-Olds[9], William F. Pfender[5,15], Thomas E. Juenger[3], Todd C. Mockler[2] and John P. Vogel[1,*]

[1]*USDA-ARS Western Regional Research Center, 800 Buchanan St., Albany, CA 94710, USA,*
[2]*Donald Danforth Plant Science Center, 975 North Warson Road, Saint Louis, MO 63132, USA,*
[3]*University of Texas at Austin, Department of Integrative Biology, 1 University Station, Austin, TX 78712, USA,*
[4]*DOE Joint Genome Institute, 2800 Mitchell Dr, Walnut Creek, CA 94598, USA,*
[5]*USDA-ARS, 3450 SW Campus Way, Corvallis, OR 97331, USA,*
[6]*University of Minnesota, 1991 Upper Buford Circle, St. Paul, MN 55108, USA,*
[7]*University of California, 1 Shields Ave, Davis, CA 95616, USA,*
[8]*University of Massachusetts, 240 Thatcher Rd., Amherst, MA 01003, USA,*
[9]*Duke Biology Box 90338, Durham, NC 27708, USA,*
[10]*Waksman Institute, Rutgers University, 190 Frelinghuysen Rd., Piscataway, NJ 08854, USA,*
[11]*Universidad de Jaén, Paraje Las Lagunillas s/n, Jaén 23071, Spain,*
[12]*USDA-ARS, Plant Science Research Unit, 1991 Upper Buford Circle, St. Paul, MN 55108, USA,*
[13]*Sabanci University, Orhanli, 34956 Tuzla-Istanbul, Turkey,*
[14]*Namik Kemal University, Department of Field Crops, 59030, Tekirdag, and*
[15]*Oregon State University, 2082 Cordley Hall, Corvallis, OR 97331, USA*

Received 7 November 2013; revised 20 May 2014; accepted 23 May 2014; published online 31 May 2014.
*For correspondence (e-mail brachypodium@gmail.com).

SUMMARY

Brachypodium distachyon **is small annual grass that has been adopted as a model for the grasses. Its small genome, high-quality reference genome, large germplasm collection, and selfing nature make it an excellent subject for studies of natural variation. We sequenced six divergent lines to identify a comprehensive set of polymorphisms and analyze their distribution and concordance with gene expression. Multiple methods and controls were utilized to identify polymorphisms and validate their quality. mRNA-Seq experiments under control and simulated drought-stress conditions, identified 300 genes with a genotype-dependent treatment response. We showed that large-scale sequence variants had extremely high concordance with altered expression of hundreds of genes, including many with genotype-dependent treatment responses. We generated a deep mRNA-Seq dataset for the most divergent line and created a *de novo* transcriptome assembly. This led to the discovery of >2400 previously unannotated transcripts and hundreds of genes not present in the reference genome. We built a public database for visualization and investigation of sequence variants among these widely used inbred lines.**

Keywords: ***Brachypodium distachyon*****, natural diversity, genome sequencing, transcriptome, drought.**

INTRODUCTION

Natural variation is a cornerstone of biology and has played a central role in understanding many basic biological processes. Early studies of natural variation necessarily focused on phenotypic diversity. Indeed, Mendel's experiments were based upon natural variation (Mendel and Bateson, 1902). Similarly, Darwin recognized natural

variation as the raw material of evolution (Darwin, 1876). Identifying genes responsible for phenotypic variation using traditional positional cloning techniques can be labor intensive, even when a reference genome is available. In addition, multiple genes with small effects often act in concert to produce the observed phenotype. Such quantitative inheritance imposes a major impediment to positional cloning of genes.

Rapid progress in sequencing technology is ushering in a new era for studies of natural variation by allowing researchers to examine variation across the entire genomes of many individuals. One approach is to sequence large numbers of individuals at low depth to conduct population genetics and/or genome-wide association studies (Huang *et al.*, 2012). While the low sequencing depth used in this approach can be sufficient to identify haplotypes or SNPs statistically associated with a particular trait it does not usually result in the identification of the causal sequence variants. Low sequencing depth is also unsuited for the identification of large and complex sequence variants (e.g. insertions, deletions, translocations and inversions).

The newest sequencing technologies allow deep sequencing of many individuals enabling unprecedented resolution of natural variation. Such studies in *Arabidopsis thaliana* (Ossowski *et al.*, 2008; Cao *et al.*, 2011), and Oryza *sativa* (Xu *et al.*, 2012) identified most small genetic variants. Therefore, the causal sequence polymorphisms for many phenotypic differences were captured. In Arabidopsis, researchers have identified 14.9 Mb of sequence deleted in one or more accession (Gan *et al.*, 2011). The authors used transcriptome analysis to observe that 45% of genes with >100-fold expression differences were associated with structural variants. Transcriptome data constitute an important, independent data set that can be used to form hypotheses about the role of specific sequence variants on gene expression (Joosen *et al.*, 2009).

Brachypodium distachyon is an annual pooid grass that occurs naturally in a circum-Mediterranean region and extending from western Europe to central Asia. It grows in pastures and open woodlands that range from very dry to moderately moist habitats (Garvin *et al.*, 2008). *B. distachyon* was first proposed as a model plant due to its small size, simple growth requirements, rapid generation time, small genome (272 Mb) and evolutionary placement between rice and wheat (Draper *et al.*, 2001). Subsequently, a wealth of experimental and genomic resources has been developed, including highly efficient *Agrobacterium*-mediated transformation, a large germplasm collection, bacterial artificial chromosome (BAC) libraries, expressed sequence tags, crossing methods and T-DNA mutants (reviewed in Brkljacic *et al.*, 2011). In addition, a high-quality Sanger-based draft genome sequence is available (IBI, 2010).

Drought has a major impact on global food production (Boyer, 1982) and this impact is increasing as precipitation patterns change in response to climate change (Hoerling *et al.*, 2011). The proposed production of biomass crops on marginal lands underscores the need to develop drought tolerant crops. Two recent studies have identified considerable phenotypic variation in *B. distachyon* water relations under field (Manzaneda *et al.*, 2012) and induced drought conditions (Luo *et al.*, 2011) highlighting the utility of *B. distachyon* as a model for drought stress.

Brachypodium distachyon is particularly attractive system for the study of natural variation as it is naturally inbred, as revealed by an extremely high level of homozygosity in wild plants, simplifying the development of inbred lines (Vogel *et al.*, 2009). Furthermore, a large collection of geographically diverse natural accessions has been established (Filiz *et al.*, 2009; Vogel *et al.*, 2009; Mur *et al.*, 2011). Many of these lines were genetically characterized using simple sequence repeat (SSR) markers, which demonstrated considerable genetic variation, although the haplotype structure is unknown (Vogel *et al.*, 2009; Mur *et al.*, 2011). Genotypic variation in these lines is paralleled with extensive phenotypic variation in traits such as disease resistance, plant stature, flowering time, drought tolerance, biomass and cell wall composition (Vogel *et al.*, 2009; Luo *et al.*, 2011; Cui *et al.*, 2012; Tyler *et al.*, 2014).

In this study we used deep sequencing to determine whole-genome sequence variation in seven diverse *B. distachyon* lines. The data and analysis presented here dramatically enable the study of the genetic basis of phenotypic diversity in these inbred lines. We present a detailed analysis of shared and unique variation at the chromosomal scale that was previously lacking. Transcriptome analysis of the most divergent line, Bd1-1, reveals genes not present in the reference genome as well as previously unannotated genes. In addition, we have characterized genome-wide gene expression in response to water deficit and examined the effect of sequence variants on gene expression under these conditions.

RESULTS

Selection of lines for resequencing

We selected lines that were highly divergent based on previous SSR marker analysis aiming to estimate diversity within the species. The line Bd1-1 from Turkey is the most divergent based on SSR markers and represents a clade of phenotypically distinct late-flowering lines (Vogel *et al.*, 2009). We also included a line from Spain, Bd30-1, that fell between the two main clades based on a subsequent analysis with many of the same SSR markers (J.P. Vogel, unpublished). In addition to maximizing diversity, we wanted to produce sequences that would be immediately useful to the community. Thus, we included two lines, Bd21-3 and Bd3-1, that were more closely related to the reference genome based on SSR markers. Bd21-3 is the

line favored for genetic transformation, and for which over 20 000 T-DNA lines have been created (Vogel and Hill, 2008; Bragg *et al.*, 2012; http://brachypodium.pw.usda.gov/TDNA/). In addition, the Bd21-3 line originates from the same collection site in Iraq as the Bd21 reference line, allowing insight into genetic variation between plants growing next to one another in the wild. As a control, we also sequenced Bd21. Recombinant inbred lines (RILs) from crosses between Bd21 and Bd3-1 have been established and genotyped, making the extensive genotypic analysis of Bd3-1 immediately useful for quantitative trait locus (QTL) mapping and positional cloning efforts (Cui *et al.*, 2012). In fact, the resequencing data have already been used to fine map a *B. distachyon* gene that confers resistance to barley stripe mosaic virus (Cui *et al.*, 2012). Lines, Koz-3 and BdTR12c from Turkey, were selected to sample other clades of the SSR tree and for their trichome, seed detachment and growth habit phenotypes (Tyler *et al.*, 2014).

Generation and alignment of Illumina sequence data to the reference genome

We generated between 135 and 253 million 75-bp Illumina paired-end reads per line (Figure S1). In addition, we produced 449 million 100-bp Illumina paired-end reads from Bd21 (IBI, 2010). Ninety-eight percent of the Bd21 reads mapped to the reference genome with 98% of read pairs mapping with proper distance and orientation. This covered 99.47% of the 272 Mb reference genome to a depth ≥3 reads. Ninety-five percent of the 0.53% of the reference genome covered by <3 reads at least partly consisted of ambiguous nucleotides filling known gaps in the reference genome sequence. Thus, excluding known gaps in the reference, >99.97% of the reference genome was covered by ≥3 reads. We sub-sampled Bd21 sequence data to estimate the sequencing depth required to obtain >99% coverage (≥3 reads) of the reference genome and found that 34-fold covered 99.4% (Figure S2). We sequenced the six divergent lines to a depth of 34–58-fold. In these lines, 93–96% of reads mapped to the reference genome and 85–96% of those reads mapped as proper pairs (Table 1). This covered 92.6–96.8% of the reference genome (depth ≥3 reads), excluding known gaps. The 3–6% lower coverage in these lines is presumably due to sequence differences relative to the reference.

SNP predictions and quality control

We created a conservative SNP set, hereafter referenced as high-confidence SNPs, consisting of the intersection of homozygous SNPs found by two programs, MAQ and SOAP (Li and Durbin, 2009; Li *et al.*, 2009) (Figure 1a and Table 1). The high-confidence SNP set consists of 4.5 million SNPs among the six lines or 2 485 097 unique genomic positions harboring a SNP in one or more lines (non-redundant set) (Figure S3). Extensive quality control using a set of known SNPs (IBI, 2010) and resequencing of the same line, Bd21, sequenced for the reference genome revealed that we had very low false-positive and false-negative error rates (Appendix 2).

As both SOAP and MAQ rely on read mapping to identify SNPs they underestimate diversity in highly divergent regions where there are multiple mismatches in each Illumina read. Therefore, to identify SNPs missed by SOAP and MAQ, we used a program, IMR/DENOM (Gan *et al.*, 2011), which utilizes local *de novo* assemblies and iterative read mapping to the reference sequence to identify SNPs (Appendix 2, Figure S4 and Table S4). IMR/DENOM identified 529 699 SNPs that were not found in the MAQ or SOAP pipelines. These IMR/DENOM-specific SNPs were twice as likely to be within 10 bp of another SNP when compared with the full non-redundant union of SNPs produced by all pipelines. This demonstrates the ability of IMR/DENOM to identify SNPs in highly divergent genomic regions. Importantly, IMR/DENOM-specific SNPs were highly accurate as indicated by the fact that 24 of 24 IMR/DENOM-specific SNPs examined by Sanger sequencing were validated. There were 3 803 592 SNPs in the non-redundant union of the MAQ, SOAP and IMR/DENOM SNP

Table 1 Read, alignment, coverage, and SNP metrics

Accession	Depth	Reads (millions)	% map[a]	% pairs[b]	MAQ SNPs	SOAP SNPs	MAQ SOAP overlap	IMR-DENOM SNPs	Bp per SNP
Bd21-3	38	153	96	94	476 849	493 487	421 607	539 159	504–570
Bd3-1	34	135	95	95	494 416	566 538	448 597	637 384	427–550
Koz-3	46	162	95	96	753 157	726 025	635 634	855 155	318–375
BdTR12 c	51	170	94	96	795 505	760 711	662 861	901 386	302–358
Bd30-1	55	218	94	95	1 227 710	1 159 266	1 050 874	1 341 737	203–235
Bd1-1	58	253	93	85	1 546 440	1 418 299	1 297 260	1 668 191	163–192
Bd21	160	449	98	98	7765	28 973	6778	4552	9.4–59.8 kb
Average[c]	47	182	94.5	93.5	882 346	854 054	752 806	990 502	

[a]Percent of reads mapped.
[b]Percent of reads mapped as proper pairs.
[c]Average excluding Bd21.

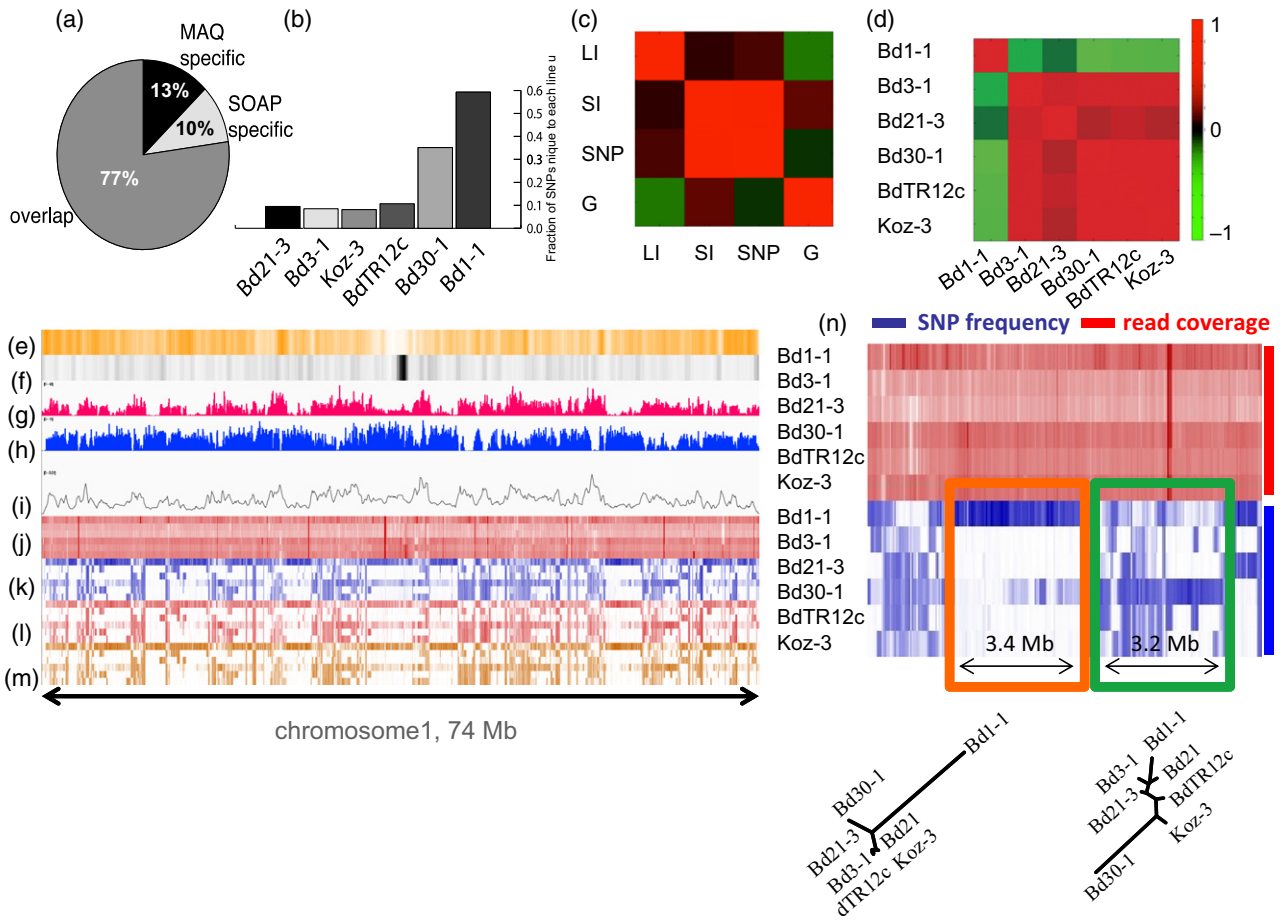

Figure 1. Genome-wide variant predictions and distribution of diversity.
(a) Overlap of MAQ and SOAP SNP predictions.
(b) Fraction of unique SNPs in each line.
(c) Heatmap of correlation coefficients among BreakDancer large indel (LI), MAQ small indel (SI), high-confidence SNP and gene densities (g) within 10 kb windows in Bd30-1 (similar trends were observed for all other lines, Figure S8). As shown by the scale in (d), red indicates positive correlation, and green indicates negative correlation.
(d) Heatmap of correlation of SNP density between lines.
(e–h) Heatmap tracks running the length of chromosome 1. Tracks display (e) gene density, (f) transposable element density, (g) SNP density of the five non-Bd1-1 lines, and (h) SNP density for Bd1-1.
(i) Graph of nucleotide diversity (250 kb sliding windows).
(j–m) Heatmap tracks plotted consecutively for Bd1-1, Bd21-3, Bd3-1, Bd30-1, Koz-3, and BdTR12c, displaying (j) read coverage, (k) SNPs, (l) deletions, and (m) insertions. Read coverage is relatively even while polymorphism density is not. Only high-confidence SNPs were used in this analysis. Note that polymorphism density in Bd1-1 is often anti-correlated with polymorphism density in the other five lines.
(n) Phylogenetic trees at the bottom of the figure show the relationship of a 3.4 Mb segment containing a high density of Bd1-1 unique SNPs and a 3.2 Mb segment containing higher levels of shared SNPs between Bd1-1 and the other lines.

sets of the six sequenced lines. This equates to one difference among the six divergent lines every 72 bp. Even though this set may have a higher false-positive error rate than the high-confidence SNP set, it is still useful for many studies.

Genome-wide indels and sequence variant predictions

We used MAQ to identify 586 206 small indels with respect to the reference genome (61 582–163 776 homozygous indels per line) (Table S3). Only 3230 homozygous indels and 696 heterozygous indels were predicted for Bd21, indicating that our pipeline has a low false-positive rate. We tested 47 homozygous indels, ranging from 7–32 bp, using the polymerase chain reaction (PCR) (Figure S5). Indels were confirmed at 38 of the 47 loci (81%). While the indels tested were predicted to be specific to Koz-3 or BdTR12c, for 10 of 33 loci both Koz-3 and BdTR12c showed the polymorphism (Table S1). Thus, indels have a high incidence of unpredicted variation.

We used BreakDancerMax-0.0.1r59 to identify larger structural variants (SV) based on read pairs that map at unexpected distances or orientations (Chen *et al.*, 2009).

We identified 2237–8413 SVs per line including larger insertions and deletions, inversions and intra- and inter-chromosomal translocations. In contrast, only 681 SVs were predicted for Bd21 and they were disproportionally (146 out of the 681) localized to scaffold six, which contains mitochondrial sequences and is not assigned to a chromosome. As alternative overlapping SVs were present in the BreakDancer output, we merged overlapping features to produce a set of 5065 non-redundant large indels. To determine the accuracy of the BreakDancer predictions, we used PCR to test 48 predicted deletions. All 24 predicted deletions between 129 and 947 bp were confirmed. In contrast, only 13 of 21 predicted deletions between 1500–5179 bp were confirmed (the reference allele was present in the unconfirmed sites) (Figure S6). We genotyped 11 additional non-sequenced lines with eight of the validated deletion markers (deletion sizes between 129 and 947 bp) and found these polymorphisms distributed broadly among *B. distachyon* populations (Figure S7). While all selected deletions were predicted to be unique to one line, eight out of nine were found in other resequenced lines, with an average of 2.4 lines containing the deletion. Thus, BreakDancerMax fails to detect a significant percentage of large deletions. As expected, the reference line did not contain any of the tested deletions. We used the Pindel program (Ye *et al.*, 2009) to create local assemblies of split-reads for the indels predicted by BreakDancer to identify the precise break points and the inserted or deleted sequence. In total, 3199 unique SVs predicted by BreakDancer were confirmed and assigned precise breakpoints and sequence information (Table S2). Pindel also confirmed 24 unique SVs in Bd21 resequencing data. Only six percent of small Pindel deletions and 0.6 percent of small insertions in the divergent lines were previously identified by MAQ (Tables S3 and S5).

To obtain the largest non-redundant set of indels, we used IMR/DENOM to align *de novo* assembled contigs against the Bd21 reference. In total, IMR/DENOM identified 1 026 556 small indels corresponding to 4.6 Mb of sequence and 14 397 large indels corresponding to 14.6 Mb among the divergent lines, a total of 19.2 Mb sequence inserted or deleted relative to the reference line (Tables S6 and S7, Appendix 2). To test the quality of the IMR/DENOM-unique indels we examined 10 loci with PCR. Eight of the 10 putative indels were validated indicating that most IMR/DENOM-unique indels are correct.

For high coverage data sets, read depth (RD) is a sensitive tool to infer copy number variations (e.g. deletions) (Abyzov *et al.*, 2011). Aligning reads from each divergent line to the Bd21 reference genome we identified 10–21 Mb of DNA covered by <3 reads, hereafter called Depth-under-3, per line. In contrast, only 1.6 Mb of low coverage sequence was obtained when 34× Bd21 sequence was used and most of that sequence was in known gaps in the reference genome. In addition to lack of depth, another signature of an SV is the presence of regions that lack new alignment starts (non-starters). Reads that should start aligning near the edge of a breakpoint will fail to align as the majority of the read aligns to another location or in unknown inserted sequence. We extended these results to the identification of duplications by segmenting genomic regions >250 bp with an observed-to-expected coverage ratio of between 1.4 and 3.1, and then selecting regions that also have high-quality heterozygous SNPs (Ossowski *et al.*, 2008). Such regions correspond to duplications in the divergent lines whose read alignments are collapsed onto the single copy present in the reference sequence (Figure 6e). As expected, no such regions were identified in Bd21. In contrast, 2–6.3 Mb of putatively duplicated sequence in the divergent lines were identified.

Genome-wide distribution of polymorphisms in the divergent lines

The density of polymorphisms of all classes was positively correlated in windows of 10 kb (Figures 1c and S8). SNP and small indel density were strongly positively correlated (Pearson's correlation = 0.75, *P*-value < 0.001). Density of BreakDancer SVs were negatively correlated with gene density (Pearson's correlation = −0.11) (Figure 1c). The density of SNPs between Bd1-1 and the reference genome versus the number of SNPs between the other resequenced lines and the reference were strongly negatively correlated along the genome (Figure 1d, Pearson's correlation = −0.31 to −0.67; 10 kb windows, *P*-value < 0.001). SNP density was otherwise strongly positively correlated among the other lines (*P*-value < 0.001). Density of small and large indels and the correlation among lines was comparable with that found for SNP density (Figures 1k–m and S9). Visualization of read coverage, SNPs, small insertions and deletions using Bd21 genome coordinates shows that the negative correlation between Bd1-1 and the remaining strains is due to genomic segments with a prevalence of variants unique to Bd1-1 interspersed with genomic segments in which diversity is shared between Bd1-1 and all lines (Figure 1e–m,n). To characterize the size the Bd1-1 unique segments and their coverage of the reference sequence, we first divided the genomes into 10 kb segments and selected windows with an above-median density of Bd1-1 unique SNPs. Segments within 100 kb were merged together, including the intervening sequence. This resulted in 384 genomic segments covering 57% of the genome with a mean size of 400 kb and a maximum size of 4.9 Mb. When polymorphisms are arranged using one of the non-Bd1-1 resequenced lines as a pseudo-reference, non-identical, but similar alternating patterns were observed. Thus, a common set of regions distinguishes Bd1-1 from all other lines in this study. Nonetheless, the high-confidence non-redundant set of SNPs from all lines

is well distributed throughout the genome (Figures S10 and S11).

For population analyses we only examined variants from sites with at least five supporting reads with a mapping quality >29 and normalized for sites lacking sufficient reads when calculating nucleotide diversity. The overall relationship between the lines is consistent with the genetic distances inferred by SSR analysis (Figure S12). As expected, Bd1-1 is the most divergent line. However, our high-resolution data reveal different relationships between smaller genome segments (Figure 1n). In genome segments containing low Bd1-1 unique variation and high nucleotide diversity, the most genetically divergent line was Bd30-1, the most geographically distant line. In contrast, in genome segments containing high Bd1-1 unique variation, Bd1-1, the latest flowering line, was the most genetically divergent as it is when the entire genome is considered (Figure 1n). Interestingly, although the reference line, Bd21, was collected at the same time and location as Bd21-3, there is extensive genetic variation between the two derived inbred lines. Indeed, while these lines are the most similar overall, in some genome regions they are the most divergent. Mean nucleotide diversity was estimated to be 0.0042 (Figure S12). The distribution of nucleotide diversity largely reflected the SNP density of the five non-Bd1-1 lines (Figure 1i), which is consistent with the large number of unique SNPs in Bd1-1 and their negatively correlated density with respect to the other lines.

Annotation and distribution of variants in genes

Using the Bd21 reference gene annotation as a guide, our high-confidence non-redundant SNP set includes SNPs in 152 920 CDSs, 48 470 UTRs, 598 splice junctions, 309 860 introns and 1 973 249 in intergenic regions (Figure 2a and Tables 2 and S8). Small indels identified by MAQ show a similar distribution among gene features with the exception that fewer indels fell into coding regions and a higher percentage fell in introns (Figure 2b). The distribution of indels within coding regions was biased towards triplet multiples, consistent with selection against disruptions in reading frames (Figure 2c) (Schneeberger *et al.*, 2011). In contrast, indels outside coding regions did not show a similar bias (Figure 2d). Within genes, the distribution of putative large-effect variants was biased toward the beginning and ends of genes as observed in other studies (Figure S13) (Ossowski *et al.*, 2008). The genes disrupted by large-effect polymorphisms are biased towards particular gene families with roles in response to biotic and abiotic stresses (Figure 2e), also consistent with previous studies (Cao *et al.*, 2011).

Line-specific genomes

Identification of polymorphisms by comparing sequences to a single reference genome is insufficient to fully understand complex sequence variants, particularly sequences absent from the reference genome. Although generating true high-quality *de-novo* assemblies necessary to fully capture polymorphisms in the resequenced genomes is beyond the scope of the current project, we used IMR/DENOM to generate genomes for each line by incorporating both large and small polymorphisms into the reference sequence. We transferred annotated Bd21 transcript coding sequences to the line-specific genomes (Figure 3a). A small and variable number of gene models failed to map to the line-specific genomes, consistent with overall divergence of the lines. Re-mapping of the genomic reads to the synthetic genome assemblies for the divergent lines resulted in 1–2.3 million additional

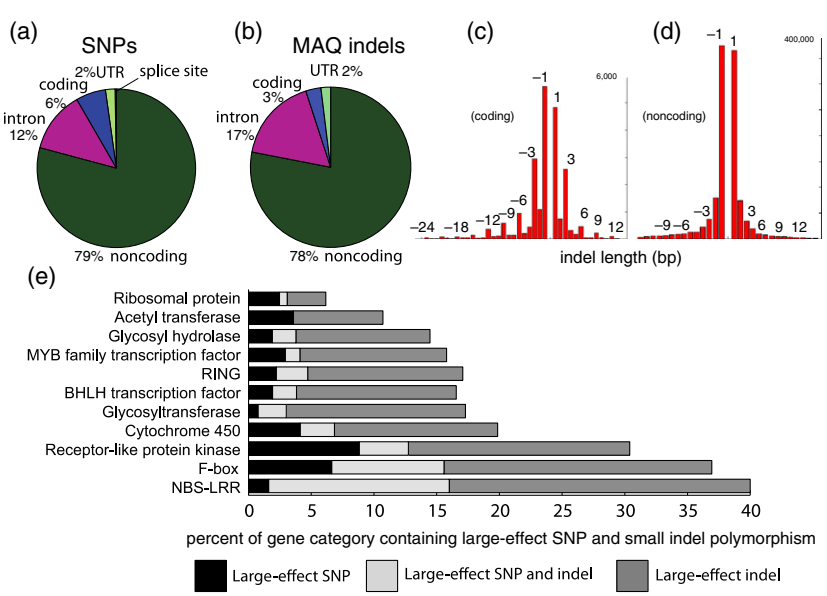

Figure 2. Annotation and distribution of variants in genes.
(a) Distribution high-confidence SNPs among gene features.
(b) A similar distribution was observed for MAQ small indels with the exception that fewer indels fell in coding regions and a larger percent fell in introns.
(c, d) Indels in coding regions (c) had a length bias in multiples of three, whereas (d) indels in non-coding regions did not. Numbers above the graphs indicate the number of bases deleted or inserted.
(e) Percentage of genes in select gene classes harboring SNPs causing nonsense changes or altering splice junctions and small indels. Note that stable gene families like ribosomal proteins have many fewer SNPs than highly variable gene families such as NBS-LRRs.

Table 2 Distribution of 2 485 097 high-confidence non-redundant SNPs

	Bd21-3	Bd3-1	Koz-3	BdTR12c	Bd30-1	Bd1-1	Frequency[a]
Coding	27 430	28 612	43 245	43 261	67 625	83 695	0.73–2.2
Intron	63 586	64 496	83 055	84 843	134 767	169 547	2.13–5.67
UTR	9958	10 348	13 063	12 911	21 208	27 026	0.56–1.53
Intergenic	320 633	345 141	496 271	521 846	827 274	1 016 992	1.72–5.44
Total	421 607	448 597	635 634	662 861	1 050 874	1 297 260	

[a]Frequency range per 1000 bp of total feature length.

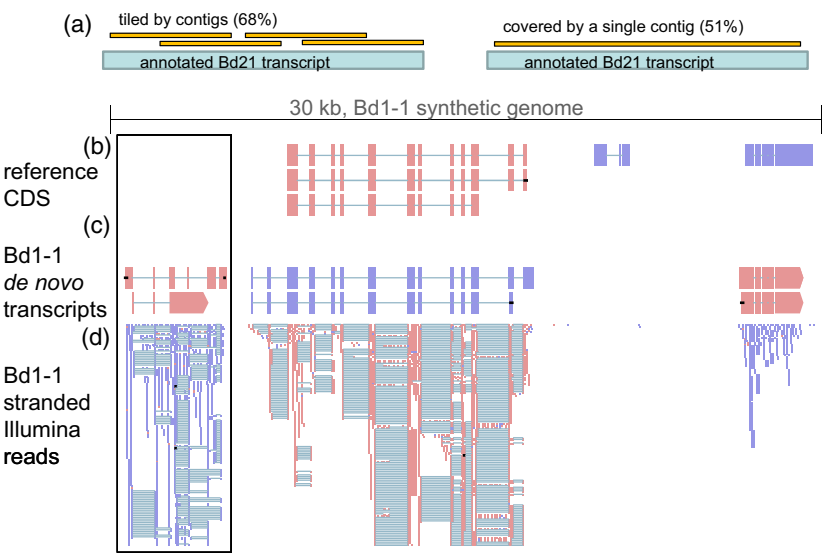

Figure 3. *De novo* Rnnotator assembly of the Bd1-1 transcriptome.
(a) Fifty-one percent of annotated Bd21 transcripts were covered by a single contig and 68% were completely tiled.
(b–d) IMR/DENOM was used to generate a genome assembly for Bd1-1 from genomic reads to which we mapped (b) annotated coding sequence for the Bd21 reference strain and (c) *de novo* contigs produced by Rnnotator. Rnnotator contigs with no equivalent in the Bd21 reference annotation are boxed. (d) TopHat alignment of a subset of 150 million stranded Bd1-1 RNA-Seq reads used to make the contigs.

mapped reads (7.6–18.0% of previously unmapped reads) and an additional 1.4–3.3 million reads that mapped as pairs with expected insert distance, depending on the line (Figures S14–S17).

Bd1-1 transcriptome assembly

Since variant discovery algorithms poorly capture sequence not found in the reference, we generated 632 million stranded RNA-Seq reads for the most divergent line, Bd1-1, and used RNNOTATOR software to assemble these reads into a *de novo* transcriptome (Martin *et al.*, 2010). The N50 of the assembly was 1021 bp with 21 916 contigs longer than 1 kb. A slightly higher percentage of Bd1-1 contigs mapped to the Bd1-1 synthetic genome as compared with the Bd21 reference, nonetheless >99% of contigs larger than 1 kb mapped in both cases. The *de novo* transcriptome had at least partial coverage for 79.5% of the Bd21 gene models mapped to the Bd1-1 synthetic genome. When compared with the Bd21 reference genome, 65% of previously annotated transcripts were completely tiled and 55% of annotated Bd21 genes were completely covered by at least one unbroken *de novo* transcript. Interestingly, 2479 contigs larger than 1 kb mapped to locations in the Bd21 reference genome that did not contain annotated genes (Figure 3b,c). When mapped to the Bd1-1 synthetic genome, these contigs had excellent support from the underlying Bd1-1 mRNA-Seq reads (Figures 3d and S18). Furthermore, 86% of the contigs not found in the reference annotation also had support from publicly available Bd21 mRNA-Seq data (Davidson *et al.*, 2012) (more than 10 reads, Bd21 mRNA-Seq mapped to Bd1-1; Figure S19). Thus, these contigs seem to represent bona-fide genes that were missed in the original annotation. A small number of these contigs involved the addition of >1 kb of sequence to beginning or end of an existing annotated transcript so they are not all entirely new genes (Figure S19).

Newly identified transcripts were distributed throughout the genome similar to Bd21 annotated transcripts (Figure S20). We annotated these contigs by assigning Gene Ontology (GO) terms through BLASTx and Interpro scan and by merging annotations through Blast2GO (Figure S21). We filtered the unannotated transcripts for sequences with similarity to transposon-associated protein sequences and putative unspliced pre-mRNA transcripts to generate a subset of 2249 transcripts associated with >1711 loci. We merged previously annotated Bd21 coding sequences with the new transcripts identified by RNNOTATOR to create a gene annotation set of 33 626 transcripts as a resource for future studies. A small set of 153 contigs >1 kb did not map to Bd1-1. These unmapped

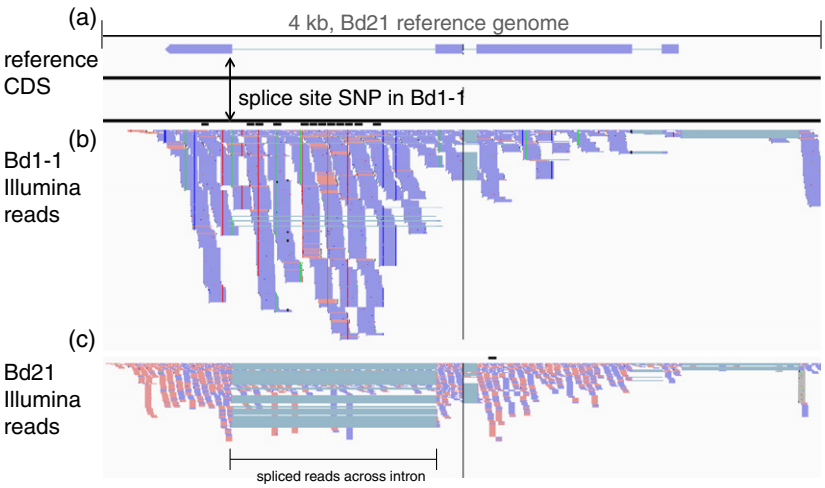

Figure 4. SNPs in splice junctions alter mRNA structure and predicted protein sequence.Two-hundred and 86 of the high-confidence Bd1-1 SNPs were predicted to modify RNA transcript splice sites.
(a) Bd21 coding sequence mapped to the Bd21 reference genome.
(b) Alignment of Bd1-1 Illumina reads to Bd21 reference genome. Vertical colored lines indicate variant bases (SNPs). The presence of the SNP in the splice junction (arrow) is associated with very low levels of splicing (intron retention) for this transcript. Transcript is read from the reverse strand and the SNP changes the acceptor splice site di-nucleotide from AG to AA. An alternative downstream AG acceptor site is used instead at low frequency in Bd1-1 due to the loss of upstream acceptor site. Use of the downstream acceptor site was not observed in Bd21. Alternative splice sites are used at low frequency in Bd1-1.
(c) Alignment of Bd21 Illumina reads to Bd21 reference genome. Light blue lines indicate the splicing of reads across introns.

contigs appear to be good assemblies because many had high E-values and sequence similarity when translated and compared with annotated proteins from rice, wheat and other grasses (Figures S22 and S23). Thus, these unmapped contigs likely contain transcribed sequences that are completely lacking from the reference genome and not captured by the IMR/DENOM variant pipeline. Unmapped sequences were enriched in kinase activity (Figures S24 and S25). Filtering these sequences for transposon and pre-mRNA transcripts generated a set of 132 transcripts associated with 109 loci. Sequences and annotation of the 153 unmapped contigs >1 kb, as well as the full, unfiltered set of 3721 unmapped contigs, are supplied in downloadable files on Brachybase.

Point mutations can activate or deactivate transcript splice sites, and thus influence the incorporation of exons into a mature messenger RNA. In this way a point mutation can have far larger effects than just the change of a single amino acid. In total, 286 of the high-confidence SNPs predicted from genomic DNA for Bd1-1 were predicted to modify RNA transcript splice sites. We used a subset of 300 million mRNA-Seq reads for Bd1-1 generated in this study to compare against 300 million publicly available mRNA-Seq reads for the reference line (Davidson *et al.*, 2012) to identify altered mRNA transcript structures resulting from SNPs in splice sites (Figure S26). Figure 4 shows an example of a SNP mutation in a splice site that is associated with increased intron retention for this transcript in Bd1-1 (Figure 4b) compared with the Bd21 mRNA-Seq data (Figure 4c). In addition, alternative low-abundance splice forms were observed.

Genome-wide expression responses to water deficit

To begin assessment of the role of genome-wide sequence variants in phenotypic diversity, we performed an mRNA-Seq experiment with the six resequenced lines and Bd21. We grew replicates of the lines in soil under control and simulated progressive drought-stress conditions in the greenhouse, and harvested tissue for RNA at the end of the treatment period. Our mRNA-Seq protocol specifically targeted the 3′ ends of mRNA transcripts (Meyer *et al.*, 2011). The experiment produced 235 million raw reads, of which 121 million passed quality control and 80.3 million mapped to the Bd21 reference. As described in Appendix 1, genotype explained some variation in mapping efficiency, but did not introduce significant bias in gene counts. This mapping represented 20 721 unique transcripts across all samples. However, 4621 of these transcripts were extremely rare (an average read count less than one among biological replicates). Summing between the two growth conditions and excluding these rare transcripts, we detected between 15 726 (in Bd30-1) and 16 624 (in Bd3-1) unique transcripts in the lines.

We used factorial analysis of variance (ANOVA) to identify constitutive and soil water deficit-induced patterns of gene expression among the lines using a controlled soil dry-down experiment (Data S1), and then identified genomic variants that might affect gene expression differences. The expression of 4072 genes showed a significant genotype effect (FDR = 0.05, see Appendix 1). Figure 5a shows the expression of a transcript, predicted to encode a cellulose synthase, which shows a typical ‘genotype’

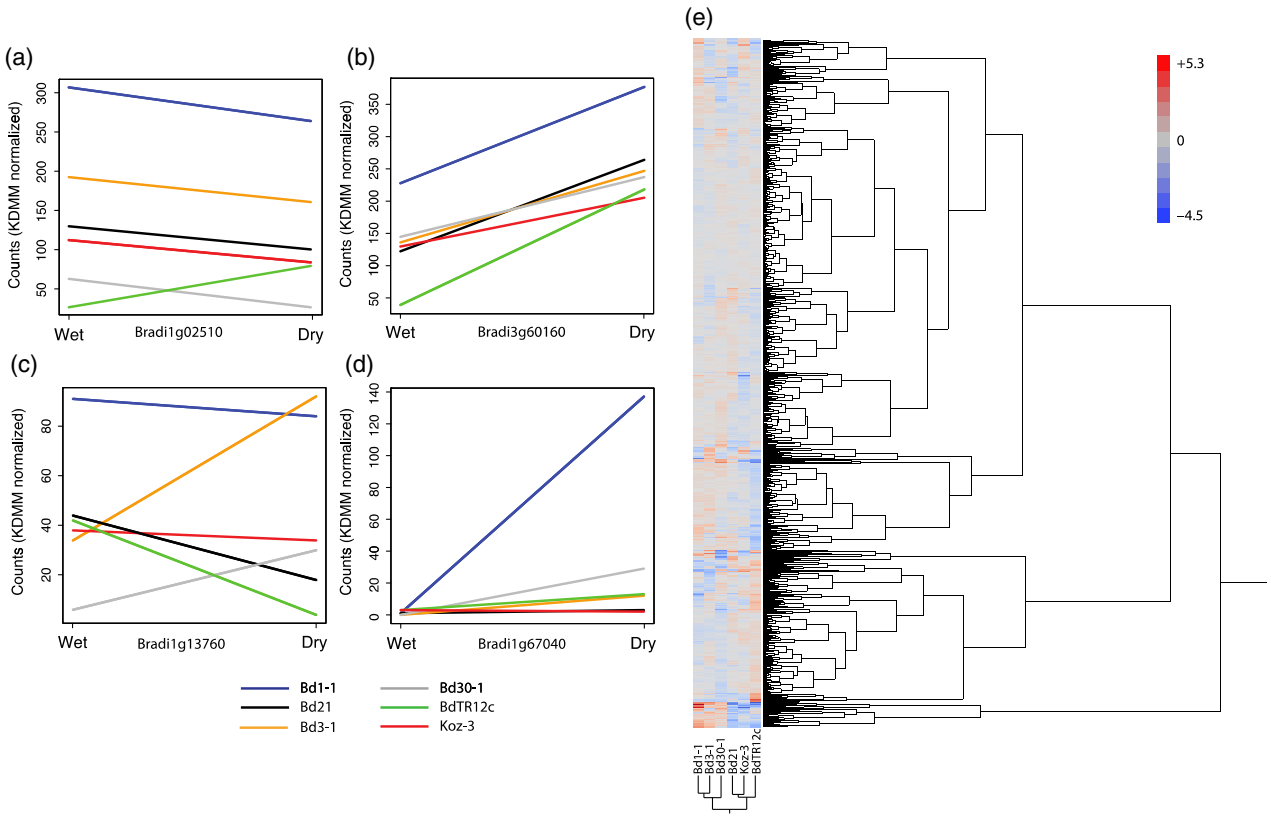

Figure 5. RNA-Seq gene expression responses of divergent lines to water deficit. Factorial ANOVA was used to identify constitutive and soil-water-deficit-induced patterns of gene expression.
(a) The expression of a transcript, predicted to encode a cellulose synthase, shows a typical 'genotype' effect in the analysis.
(b) Expression pattern for a typical 'treatment' effect transcript.
(c, d) Levels of a transcript (predicted 9-*cis*-epoxycarotenoid dioxygenase) (c) with strong rank-changing responses and (d) expression of a transcript (predicted HSP20 family member) with strong condition-dependent GxE.
(e) Cluster analysis of assayed genes. Note the dramatically different responses for some lines.

effect in the analysis. We also found that 870 genes responded significantly to our imposed soil-drying treatment (Figure 5e). Figure 5b shows a typical 'treatment' effect transcript.

In total, 300 genes show genotype-dependent treatment response (GxE) in the full-model ANOVA, indicating considerable diversity in the transcriptional response to soil drying. We used the program ELEMENT to determine whether these GxE genes were enriched for specific transcription factor binding motifs (Mockler *et al.*, 2007). The promoters of these GxE genes are enriched for the LTRE motif ($P = 0.026$), which is believed to play a role in abscisic acid (ABA)-mediated environmental signaling in Arabidopsis and cereals (Baker *et al.* 1994, White *et al.* 1994). Several interesting patterns of GxE effects can be discerned, including differential expression of transcripts with strong rank-changing responses (e.g. a gene predicted to encode a 9-*cis*-epoxycarotenoid dioxygenase protein; Figure 5c) and strongly condition-dependent responses (e.g. a predicted member of the HSP20 gene family; Figure 5d).

Large sequence variants associated with genotype-dependent expression variance

Large indels can dramatically affect gene expression (e.g. a deleted gene cannot be expressed). The location of genes with high expression variance, based on ANOVA analysis of gene expression among the divergent lines, was weakly correlated or not correlated with the location of BreakDancer indels (correlation coefficient 0.012; 10 kb windows, P-value < 0.001). In contrast, the distribution of indels was negatively correlated with genes as a whole (correlation coefficient −0.11; 10 kb windows, P-value < 0.001). Of 414 genes represented in our mRNA-Seq data set that overlapped at least 85% with a BreakDancer predicted deletion, 185 (45%) were not expressed in the line containing the deletion but were expressed in other lines. Figure 6a–d shows the expression effects of a deletion of a cluster of predicted NBS-LRR genes. IMR/DENOM deletions had a 59% (29/49) concordance with a lack of gene expression. Pindel deletions had only a 19%

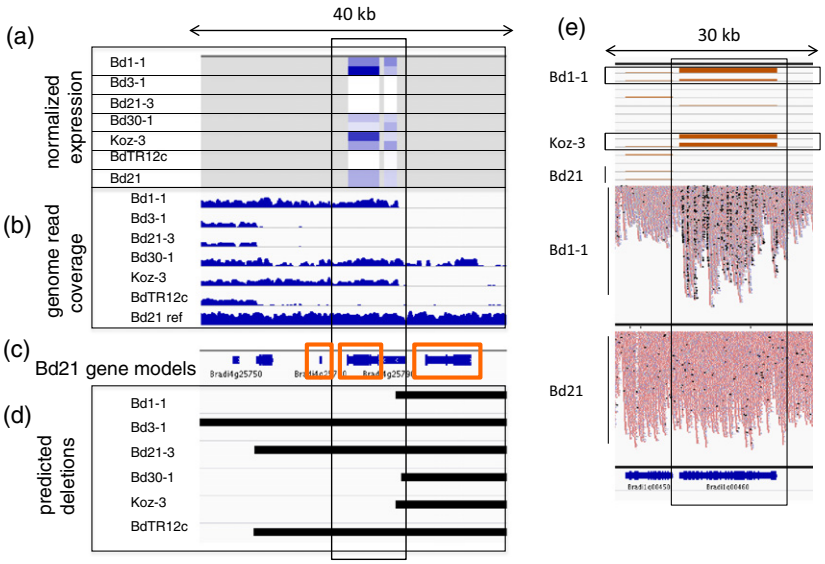

Figure 6. Functional consequences of large deletion variants on gene expression.
(a) Heatmap for gene expression in plants in dry versus watered conditions for two adjacent genes with a high variance of expression. Intensity of blue shading indicates level of expression. Gray shading indicates data not shown.
(b) Plots of genomic read coverage over the same region.
(c) Track of annotated Bd21 gene models.
(d) Large deletions, represented by black bars, predicted by BreakDancer in each line. The presence/absence of predicted deletions is consistent with the presence/absence of gene expression. Orange box in panel (c) highlights three tandem predicted genes annotated as NB-ARC domain disease resistance genes. Black box highlights two genes represented in the mRNA-Seq experiment (shown in (a)) and overlapping a deletion of the region in a subset of lines (shown in (d)).
(e) Putative transporter encoding gene Bradi1 g00460 identified as having GxE interaction for expression in the soil dry-down experiment. This sequence is duplicated in Bd1-1 and Koz-3, resulting in higher gene expression (top panel), and a peak of higher read depth with a large number of high-quality heterozygous polymorphisms (shown for Bd1-1 in middle panel, polymorphisms in black) that is not observed in the other sequenced lines (Bd21 read alignments shown in lower panel).

Table 3 Genome-wide deletion/gene expression concordance

Deletion type	% gene overlap	% lacking expression
BreakDancer	85	45
IMR/DENOM	85	59
Pindel	85	19
Nonstarter	85	97
Depth-under-3	85	86
Depth-under-3	55	45
Depth-under-3	Intersection	4.75
Depth-under-3/BreakDancer	85	91
Depth-under-3/IMR-DENOM	85	78
Depth-under-3/Pindel	85	80

(45/233) concordance with gene expression (Table 3). In contrast, 86% (156/181) of Depth-under-3 putative deletions overlapping 85% or more of a gene represented in the mRNA-Seq data set were associated with lack of expression for that gene in the line with the deletion, while still expressed in other lines. Illumina sequencing is known to suffer from under-representation and reduced quality at loci with extreme base compositions such as excessively GC rich or poor sequences (Aird *et al.*, 2011). The mean GC content of genic sequence with an overlapping Depth-under-3 putative deletion associated with lack of expression in some lines ranged from 44 to 46%, essentially the same as the average GC content of the whole genome (46%). Thus, GC content was not a major contributor to low genomic or mRNA-Seq read coverage.

Similar concordance between lack of gene expression and genomic read coverage was observed for non-starters. Of non-starters overlapping 85% of a gene, 97% (57/59) were associated with lack of expression in that line (Table 3). Genomic duplications were correlated with several fold higher gene expression in particular lines. For example, Figure 6e shows a putative transporter from the list of genes showing GxE interaction for expression in the soil dry-down experiment that has a duplicated copy in Bd1-1 and Koz-3. Transcripts from both copies in Bd1-1 and Koz-3 are collapsed onto the single copy in Bd21 contributing to higher quantified expression. The increased expression is more than additive implying that one of the duplicate copies has diverged to be more highly expressed than the *Bd21* gene. Notably, overexpression of the Arabidopsis ortholog of this gene confers drought and salt tolerance (Kim *et al.*, 2010).

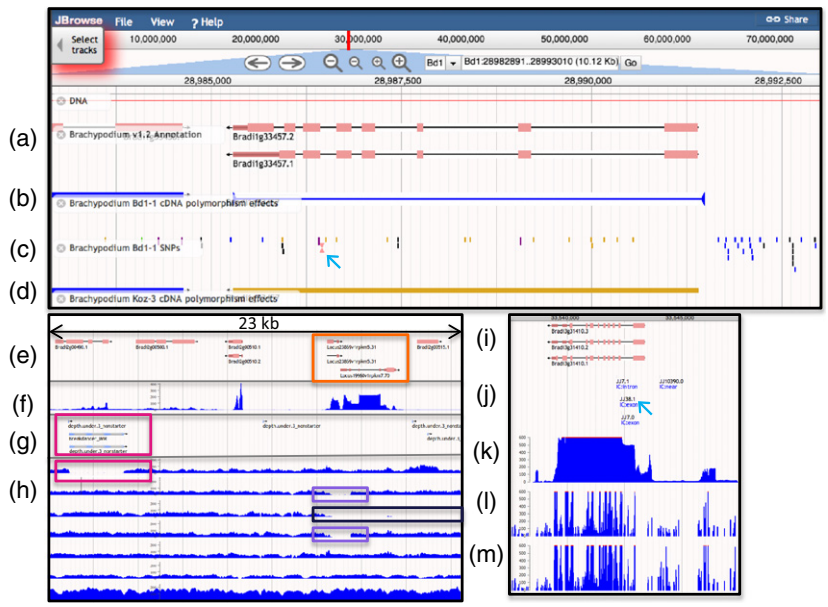

Figure 7. Screenshots from JBrowse on Brachybase.org highlighting some of the resequencing data. Thirty-six data tracks from the present study can be viewed by selecting tracks using the 'select tracks' button at the top right. The legend for the color and glyph code for all tracks can be viewed by selecting the '?Help, Track Display Details &Color Key for SNPs' button at the top right.
(a) This track shows the Munich Information Center for Protein Sequences version 1.2 annotation of the reference genome. The gene in the center, *Bradi1 g33457*, encodes a predicted protein kinase and has two predicted splice variants.
(b) This track shows the predicted effect of SNP and small indel polymorphisms on gene models. The blue bar glyph indicates that at least 20% of the predicted protein is missing in Bd1-1 due to a stop codon or frameshift.
(c) Bd1-1 SNPs color coded to indicate effect on the predicted protein. The blue arrow points to a pink hourglass glyph indicating a stop codon. At higher magnification this track shows the actual base changes and at lower magnification it shows a graph of SNPs per DNA segment.
(d) The same gene model in line Koz-3. The gold color indicates that only SNPs causing synonymous substitutions are found in this region.
(e) Merged Phytozome 8 gene annotation with new gene models from this study derived from Rnnotator Bd1-1 contigs mapped and annotated on the Bd21 reference sequence. The orange box highlights two new gene models.
(f) Bd1-1 mRNA-Seq read coverage.
(g) Consequences of large deletion and duplication variants are not included in annotation of polymorphic gene models (Figure 7b) but can be viewed as a separate track. Tracks containing non-redundant BreakDancer and IMR/DENOM SV predictions added to non-redundant Depth-under-3 and Nonstarter features as well as putative duplications are available for each line. In this example, BreakDancer/IMR-DENOM predicts a deletion overlapping Bradi2 g00490 (highlighted by the mauve box).
(h) Read coverage for each line supports the predicted deletion of the Bradi2 g00490 sequence from Koz-3, but not other lines. The read coverage tracks also reveal deletions overlapping gene models that BreakDancer and IMR/DENOM did not detect. A large 10 kb putative deletion (brown box) is present in Bd3-1 that overlaps two newly annotated genes (Locus23869v1rpkm5.3 and Locus19960v1rpkm7.70) as well as Bradi2 g00515. A smaller 1.4 kb deletion (purple box) is present in BdTR12c and Bd30-1 overlapping Locus23869v1rpkm5.3. Read coverage tracks are ordered: Koz-3, BdTR12c, Bd3-1, Bd30-1, Bd1-1, Bd21-3, Bd21.
(i) Current gene models for Bradi3 g31410, predicted to encode a trehalase identified as differentially expressed in response to our simulated drought experiments.
(j) T-DNA insertion lines are predicted to be available for this gene.
(k) Bd1-1 mRNA-Seq read coverage for this gene.
(l) Microarray expression data for Bd21 after 24 h of control treatment.
(m) Microarray expression data for Bd21 after 24 h of heat treatment. Database URL: http://jbrowse.brachypodium.org/JBrowse.html.

Online data visualization and download

Data generated in this study can be accessed through Brachypodium.org (Figure 7). SNPs and gene models can be visualized through JBrowse and are color-coded based on predicted functional annotation and quality (Skinner *et al.*, 2009). Each gene locus links to the annotation, with further links for nucleotide and protein sequence download. In addition, the high-confidence SNP set, full non-redundant SNP set, small indel predictions, BreakDancer SVs, Pindel SVs, read-depth deletion predictions, IMR/DENOM variants, IMR/DENOM line-specific genomes and annotations, transcript sequences not in the current reference gene annotation, transcripts not mapped to the reference genome, and genes showing GxE interactions in our drought study are available for download (ftp://brachypodium.org/brachypodium.org/). All sequence generated in this study is available from the sequence read archive.

DISCUSSION

We generated the equivalent of 900-fold sequence depth of the 272 Mb *B. distachyon* genome for seven highly divergent lines. Multiple methods, programs and controls were utilized to identify polymorphisms and validate their quality. Considerable effort was focused on identifying larger

polymorphisms that are difficult to detect using short-read sequencing. We used mRNA-Seq data as an independent dataset to distinguish between local deletions of genic sequence versus removal of gene sequence from the genome. The prevalence of mRNA-Seq expression for genes overlapped by PEM, SR and assembly-based large deletion variants suggests that the underlying affected sequence is present elsewhere in the genome. In contrast, RD was an effective filter for identifying gene sequence that is highly dissimilar or deleted from the genome in the sequenced lines. As a resource we incorporated polymorphisms into genome assemblies for each line.

The level of nucleotide diversity estimated in this study for *B. distachyon* is similar to that reported for the wild selfing species *Medicago truncatula* and *Arabidopsis thaliana* and the partially outcrossing ancestor of domesticated rice, *Oryza rufipogon* (Nordborg *et al.*, 2005; Caicedo *et al.*, 2007; Branca *et al.*, 2011). In contrast, the level of polymorphism is greater than that observed for domesticated soybean or rice and less than observed in the outcrossing domesticated grass, maize (Caicedo *et al.*, 2007; Gore *et al.*, 2009; Branca *et al.*, 2011). Lines Bd21-3 and Bd21 originate from the same collection site in Iraq. Despite this, these two lines are genetically very distinct. This may have arisen due to long distance seed dispersal and been maintained by the selfing nature of *B. distachyon* (Vogel *et al.*, 2009).

A complex segmental distribution of polymorphism in blocks up to 4.9 Mb was observed (Figure 1e–m). Bd1-1 (the most divergent line) polymorphism was negatively correlated with polymorphism in all other lines relative to the reference genome. There is an excess of Bd1-1-unique SNPs relative to pairwise differences in genome segments resulting in low D-statistic (Figures S27–S29). Bd1-1 is morphologically distinct and belongs to a genetically defined group of late-flowering lines (Vogel *et al.*, 2009). Further experiments are needed to determine whether the pattern of polymorphism observed is due to persistent selection or recent admixture. One possible explanation is that the presence of subgenomic segments with low D-statistic is driven by adaptive genes for flowering in that region.

We generated a *de novo* transcriptome assembly of the most divergent line, Bd1-1. The assembly identified >2400 transcripts not present in the reference gene annotation. As most are present in the reference genome and were expressed in Bd21, these are bona-fide genes missing from the current annotation. In total, 153 transcripts >1 kb did not match the reference genome and thus represent Bd1-1 unique genes. They were enriched in predicted kinase activity, polysaccharide binding, oxidative stress response and water deprivation (Figures S24 and S25). One of the unmapped transcripts is similar to MADS-box transcription factor 1 that alters flowering time (Jeon *et al.*, 2000). Thus, it may contribute to the late-flowering phenotype of the Bd1-1.

Brachypodium distachyon's range spans many climate zones suggesting that populations are locally adapted. Using 3′ mRNA-Seq to examine gene expression, we identified 870 genes that responded significantly to water stress. Two genes with conserved stress response in *B. distachyon* are of particular interest. Bradi3 g31410 is predicted to encode a trehalase and overexpression of its Arabidopsis ortholog, ATTRE1, leads to increased drought tolerance (Van Houtte *et al.*, 2013). Bradi3 g42780 is a predicted fatty acid reductase and members of this family synthesize the cuticular hydrocarbons that participate in drought avoidance strategies in plants (Shepherd and Wynne Griffiths, 2006). We found that many of the GxE genes had promoter motifs associated with ABA-mediated signaling. Sequencing additional *Brachypodium* lines will facilitate the use of eQTL-mapping approaches to identify causal variants due to distantly acting (*trans*-)features such as miRNAs, transcription factors, or other signaling components. The transcriptome analysis and data provided here, particularly for non-reference accessions, is already an enormous resource that was previously lacking.

We also observed that expression of 4072 genes varied significantly between lines and that some of these expression differences were correlated with putative deletions. In total, more than 11.3 Mb of non-redundant sequence are inserted or deleted in the six resequenced lines and hundreds of genes are predicted to be disrupted by large indels (Figures 6a–e and S17). Large deletions supported by read-depth algorithms had extremely high concordance with a lack of gene expression. Structural variation in the soybean genome localizes to clusters of biotic stress-response genes (McHale *et al.*, 2012). Although SVs identified in our study are much more widely distributed, likely due to the higher overall diversity observed, numerous identified SVs overlap clusters of NB-ARC and NBS-LRR domain-containing proteins (Figure 6a–d).

Our validated datasets demonstrate the existence of extensive natural variation in *B. distachyon.* Thus, it will be possible to apply experimental approaches based on natural variation (e.g. association and QTL mapping) to identify gene regions as well as genes that are responsible for phenotypic variation. Indeed, our SNP and indel data have already been used to map a disease resistance gene to an interval of only 26 kb (Cui *et al.*, 2012). This demonstrates the utility of our resequencing data as a foundation for studies of natural diversity in *B. distachyon.*

EXPERIMENTAL PROCEDURES

Genomic DNA sequence generation and variant analysis

High molecular weight nuclear genomic DNA was isolated as described previously (Peterson *et al.*, 2000). Illumina sequencing was carried out at the Joint Genome Institute and variants identified as described in Appendix 1.

Quantitative mRNA-Seq data generation and analysis

Plants were grown in a 2:1 mixture of Sunshine MVP soil (Sungro Horticulture, Bellevue, WA, USA): Turface MVP (Turface Athletics, Buffalo Grove, IL, USA) in the greenhouse facility at the University of Texas at Austin. Two replicates of each genotype were assigned to three randomized blocks (42 plants in total). After 10 days of cold treatment at 6°C, the plants were transferred to the greenhouse with 16-h days and watered to field capacity every other day. During the experiment, daytime temperature highs averaged 26°C and nighttime lows averaged 17°C. After 25 days of growth, all pots were brought to 75% of field capacity as estimated by pot weight. Plants were then randomly assigned, within block, to two treatments. 'Control' plants were watered daily to 75% field capacity while 'Treatment' plants experienced a progressive but controlled soil drying: day 1 70%, day 2 60%, day 3 50%, day 4 40%, day 5 30% field capacity. All Bd30-1 plants and both BdTR12c plants in block 2 flowered during the dry-down treatment; no other plants flowered before harvest. On day 5 of the dry down between 13:00 h and 13:30 h, we sampled the youngest, fully expanded leaves of the two tallest tillers for mRNA analysis. Details for RNA extraction, sequencing and analysis methods are described in Appendix 1.

Bd1-1 deep RNA sequencing and transcriptome assembly

Bd1-1 seedlings were grown as previously described (Figueroa *et al.*, 2013). cDNA libraries were prepared from three independent experiments, and each experiment consisted of a collection of eight different indexed cDNA libraries. Each cDNA library was constructed from 30 tertiary leaves collected from 3-week-old Bd1-1 seedlings. Leaves were collected and immediately frozen in liquid nitrogen. Total RNA extractions were performed as described (Filichkin and Mockler, 2012). Poly(A)+ RNA was isolated from 50 μg of total RNA using the Ambion® MicroPoly(A) Purist™ Kit (Invitrogen, https://www.lifetechnologies.com) according to the manufacturer's instructions. cDNA libraries were constructed using ScriptSeq™ v2 RNA-Seq Library Preparation Kit (Epicentre, http://www.epibio.com/) with 50 ng of poly(A)+ RNA per reaction. Ten cycles of PCR and reverse oligos corresponding to indexes (1–8 available from Epicentre) were used during cDNA library amplification. Sequencing and analysis details are described in Appendix 1.

Accession numbers

SRP010590 (*Brachypodium distachyon* Bd3-1 Project), SRP010886 (*Brachypodium distachyon* Bd21-3 Project), SRS190935 (*Brachypodium distachyon* Bd1-1 Project), SRS190910 (*Brachypodium distachyon* Bd30-1 Project), SRS190847 (*Brachypodium distachyon* BdTR12c Project), SRS190848 (*Brachypodium distachyon* Koz-3 Project), SRR1425915 (Bd1-1 transcriptome) and SRR1427166-SRR1427172 (3′ RNA-Seq data).

ACKNOWLEDGEMENTS

We thank Uffe Hellsten for help with population genetics. Genomic DNA sequencing was performed at the US Department of Energy Joint Genome Institute through the Community Sequencing Program. Deep sequencing of the Bd1-1 transcriptome was performed at the Center of Genome Research and Biocomputing, Oregon State University, Corvallis, OR. This work was supported by the Office of Science (BER), US Department of Energy, USDA NIFA, and by the USDA CRIS project 5325-21000-017-00.

SUPPORTING INFORMATION

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

Figure S1 Insert size distribution.
Figure S2 Genome coverage vs. sequence depth.
Figure S3 Saturation plot.
Figure S4 SNP quality scores.
Figure S5 Small indel validation.
Figure S6 BreakDancer variant validation.
Figure S7 Indels in unsequenced lines.
Figure S8 Variant correlations.
Figure S9 Polymorphism distribution.
Figure S10 Unique SNP distribution.
Figure S11 Total SNPs along chromosomes.
Figure S12 Diversity estimates.
Figure S13 Genic positions of large effect variants.
Figure S14 Line-specific genomes improve short-read mapping.
Figure S15 Reference genome error.
Figure S16 Bd1-1 genome alignment.
Figure S17 Bd1-1 genome alignment.
Figure S18 Bd1-1-specific transcript.
Figure S19 Extension of known genes.
Figure S20 Distribution of unannotated transcripts.
Figure S21 Annotation of new genes.
Figure S22 Annotation of unmapped transcripts.
Figure S23 BLAST annotation.
Figure S24 GO terms.
Figure S25 GO term distribution.
Figure S26 Alternate splicing.
Figure S27 Tajima's D statistic vs. SNPs.
Figure S28 Tajima's D along chromosomes.
Figure S29 Tajima's D vs. fixed SNPs.

Table S1 Results of validation of MAQ small indel predictions using PCR primers flanking the predicted polymorphic region
Table S2 BreakDancer and Pindel structural variation overlap
Table S3 MAQ small indels and Pindel structural variant overlap
Table S4 BreakDancer, Pindel and IMR/DENOM structural variant metrics.
Table S5 Percent overlap of IMR/DENOM variants with MAQ and Pindel variants
Table S6 IMR estimates of bases altered by diversity
Table S7 IMR estimates of bases altered by variants in accessions
Table S8 Functional annotation of common set SNPs in gene features

Data S1 Differentially expressed genes.

Appendix 1 Additional experimental procedure details.

Appendix 2 AQC and analysis details.

REFERENCES

Abyzov, A., Urban, A.E., Snyder, M. and Gerstein, M. (2011) CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* **21**, 974–984.

Aird, D., Ross, M.G., Chen, W.S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D.B., Nusbaum, C. and Gnirke, A. (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.* **12**, R18.

Baker, S. S., Wilhelm, K. S. and Thomashow, M. F. (1994) The 5′-region of Arabidopsis thaliana cor15a has cis-acting elements that confer cold-, drought-, and ABA-regulated gene expression. *Plant Mol. Biol.* **24**, 701–713.

Boyer, J.S. (1982) Plant productivity and environment. *Science*, **218**, 443–448.

Bragg, J.N., Wu, J., Gordon, S.P., Guttman, M.E., Thilmony, R., Lazo, G.R., Gu, Y.Q. and Vogel, J.P. (2012) Generation and characterization of the

western regional research center *Brachypodium* T-DNA insertional mutant collection. *PLoS ONE*, **7**, e41916.

Branca, A., Paape, T.D., Zhou, P. *et al.* (2011) Whole-genome nucleotide diversity, recombination, and linkage disequilibrium in the model legume *Medicago truncatula*. *Proc. Natl. Acad. Sci. U.S.A.* **108**, E864–E870.

Brkljacic, J., Grotewold, E., Scholl, S. *et al.* (2011) *Brachypodium* as a model for the grasses: today and the future. *Plant Physiol.* **157**, 3–13.

Caicedo, A.L., Williamson, S.H., Hernandez, R.D. *et al.* (2007) Genome-wide patterns of nucleotide polymorphism in domesticated rice. *PLoS Genet.* **3**, 1745–1756.

Cao, J., Schneeberger, K., Ossowski, S. *et al.* (2011) Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nat. Genet.* **43**, 956–963.

Chen, K., Wallis, J.W., McLellan, M.D. *et al.* (2009) BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods*, **6**, 677–681.

Cui, Y., Lee, M.Y., Huo, N. *et al.* (2012) Fine mapping of the Bsr1 barley stripe mosaic virus resistance gene in the model grass *Brachypodium distachyon*. *PLoS ONE*, **7**, e38333.

Darwin, C. (1876) *The Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*, 6th edn, with additions and corrections edn. London: John Murray.

Davidson, R.M., Gowda, M., Moghe, G., Lin, H., Vaillancourt, B., Shiu, S.H., Jiang, N. and Robin Buell, C. (2012) Comparative transcriptomics of three Poaceae species reveals patterns of gene expression evolution. *Plant J.* **71**, 492–502.

Draper, J., Mur, L.A.J., Jenkins, G., Ghosh-Biswas, G.C., Bablak, P., Hasterok, R. and Routledge, A.P.M. (2001) *Brachypodium distachyon*. A new model system for functional genomics in grasses. *Plant Physiol.* **127**, 1539–1555.

Figueroa, M., Alderman, S., Garvin, D.F. and Pfender, W.F. (2013) Infection of *Brachypodium distachyon* by formae speciales of *Puccinia graminis*: early infection events and host-pathogen incompatibility. *PLoS ONE*, **8**, e56857.

Filichkin, S.A. and Mockler, T.C. (2012) Unproductive alternative splicing and nonsense mRNAs: a widespread phenomenon among plant circadian clock genes. *Biol. Direct*, **7**, 20.

Filiz, E., Ozdemir, B.S., Budak, F., Vogel, J.P., Tuna, M. and Budak, H. (2009) Molecular, morphological, and cytological analysis of diverse *Brachypodium distachyon* inbred lines. *Genome*, **52**, 876–890.

Gan, X., Stegle, O., Behr, J. *et al.* (2011) Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. *Nature*, **477**, 419–423.

Garvin, D., Gu, Y.-Q., Hasterok, R., Hazen, S., Jenkins, G., Mockler, T., Mur, L. and Vogel, J. (2008) Development of genetic and genomic research resources for *Brachypodium distachyon*, a new model system for grass crop research. *Crop Sci.* **48**, S69–S84.

Gore, M.A., Chia, J.M., Elshire, R.J. *et al.* (2009) A first-generation haplotype map of maize. *Science*, **326**, 1115–1117.

Hoerling, M., Eischeid, J., Perlwitz, J., Quan, X., Zhang, T. and Pegion, P. (2011) On the increased frequency of Mediterranean drought. *J. Clim.* **25**, 2146–2161.

Huang, X., Kurata, N., Wei, X. *et al.* (2012) A map of rice genome variation reveals the origin of cultivated rice. *Nature*, **490**, 497–501.

IBI (2010) Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature*, **463**, 763–768.

Jeon, J.S., Jang, S., Lee, S. *et al.* (2000) Leafy hull sterile1 is a homeotic mutation in a rice MADS box gene affecting rice flower development. *Plant Cell*, **12**, 871–884.

Joosen, R.V., Ligterink, W., Hilhorst, H.W. and Keurentjes, J.J. (2009) Advances in genetical genomics of plants. *Curr. Genomics*, **10**, 540–549.

Kim, D.Y., Jin, J.Y., Alejandro, S., Martinoia, E. and Lee, Y. (2010) Overexpression of AtABCG36 improves drought and salt stress resistance in *Arabidopsis*. *Physiol. Plant*, **139**, 170–180.

Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li, R., Li, Y., Fang, X., Yang, H., Wang, J. and Kristiansen, K. (2009) SNP detection for massively parallel whole-genome resequencing. *Genome Res.* **19**, 1124–1132.

Luo, N., Liu, J., Yu, X. and Jiang, Y. (2011) Natural variation of drought response in *Brachypodium distachyon*. *Physiol. Plant*, **141**, 19–29.

Manzaneda, A.J., Rey, P.J., Bastida, J.M., Weiss-Lehman, C., Raskin, E. and Mitchell-Olds, T. (2012) Environmental aridity is associated with cytotype segregation and polyploidy occurrence in *Brachypodium distachyon* (Poaceae). *New Phytol.* **193**, 797–805.

Martin, J., Bruno, V.M., Fang, Z., Meng, X., Blow, M., Zhang, T., Sherlock, G., Snyder, M. and Wang, Z. (2010) Rnnotator: an automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads. *BMC Genomics*, **11**, 663.

McHale, L.K., Haun, W.J., Xu, W.W., Bhaskar, P.B., Anderson, J.E., Hyten, D.L., Gerhardt, D.J., Jeddeloh, J.A. and Stupar, R.M. (2012) Structural variants in the soybean genome localize to clusters of biotic stress-response genes. *Plant Physiol.* **159**, 1295–1308.

Mendel, G.B. and Bateson, W. (1902) *Mendel's Principles of Heredity: A Defence, with a Translation of Mendel's Original Papers on Hybridisation*, 1st edn. Cambridge, UK: Cambridge University Press.

Meyer, E., Aglyamova, G.V. and Matz, M.V. (2011) Profiling gene expression responses of coral larvae (*Acropora millepora*) to elevated temperature and settlement inducers using a novel RNA-Seq procedure. *Mol. Ecol.* **20**, 3599–3616.

Mockler, T.C., Michael, T.P., Priest, H.D., Shen, R., Sullivan, C.M., Givan, S.A., McEntee, C., Kay, S.A. and Chory, J. (2007) The DIURNAL project: DIURNAL and circadian expression profiling, model-based pattern matching, and promoter analysis. *Cold Spring Harb. Symp. Quant. Biol.* **72**, 353–363.

Mur, L.A., Allainguillaume, J., Catalan, P., Hasterok, R., Jenkins, G., Lesniewska, K., Thomas, I. and Vogel, J. (2011) Exploiting the Brachypodium Tool Box in cereal and grass research. *New Phytol.* **191**, 334–347.

Nordborg, M., Hu, T.T., Ishino, Y. *et al.* (2005) The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biol.* **3**, e196.

Ossowski, S., Schneeberger, K., Clark, R.M., Lanz, C., Warthmann, N. and Weigel, D. (2008) Sequencing of natural strains of *Arabidopsis thaliana* with short reads. *Genome Res.* **18**, 2024–2033.

Peterson, D.G., Tomkins, J.P., Frisch, D.A., Wing, R.A. and Paterson, A.H. (2000) Construction of plant bacterial artificial chromosome (BAC) libraries: an illustrated guide. *J. Agric. Genomics*, **5**, 1–100.

Schneeberger, K., Ossowski, S., Ott, F. *et al.* (2011) Reference-guided assembly of four diverse *Arabidopsis thaliana* genomes. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 10249–10254.

Shepherd, T. and Wynne Griffiths, D. (2006) The effects of stress on plant cuticular waxes. *New Phytol.* **171**, 469–499.

Skinner, M.E., Uzilov, A.V., Stein, L.D., Mungall, C.J. and Holmes, I.H. (2009) JBrowse: a next-generation genome browser. *Genome Res.* **19**, 1630–1638.

Tyler, L., Fangel, J.U., Fagerstrom, A.D., Steinwand, M.A., Raab, T.K., Willats, W.G. and Vogel, J.P. (2014) Selection and phenotypic characterization of a core collection of *Brachypodium distachyon* inbred lines. *BMC Plant Biol.* **14**, 25.

Van Houtte, H., Vandesteene, L., Lopez-Galvis, L. *et al.* (2013) Overexpression of the trehalase gene *AtTRE1* leads to increased drought stress tolerance in *Arabidopsis* and is involved in abscisic acid-induced stomatal closure. *Plant Physiol.* **161**, 1158–1171.

Vogel, J. and Hill, T. (2008) High-efficiency *Agrobacterium*-mediated transformation of *Brachypodium distachyon* inbred line Bd21-3. *Plant Cell Rep.* **27**, 471–478.

Vogel, J.P., Tuna, M., Budak, H., Huo, N., Gu, Y.Q. and Steinwand, M.A. (2009) Development of SSR markers and analysis of diversity in Turkish populations of *Brachypodium distachyon*. *BMC Plant Biol.* **9**, 88.

White, A. J., Dunn, M. A., Brown, K. and Hughes, M. (1994) Comparative analysis of genomic sequence and expression of alipid transfer protein gene family in winter barley. *J. Exp. Bot.* **45**, 1885–1892.

Xu, X., Liu, X., Ge, S. *et al.* (2012) Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nat. Biotechnol.* **30**, 105–111.

Ye, K., Schulz, M.H., Long, Q., Apweiler, R. and Ning, Z. (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, **25**, 2865–2871.