Supporting Information

Materials and Methods

Plant materials and growth conditions

We used 201 *A. thaliana* accessions obtained from the Arabidopsis Biological Resource Center (Dataset S1). These accessions were part of the 1,135 accessions sequenced by the 1001 Genomes Project for *A. thaliana* (http://1001genomes.org/). Accession Col-0 was used as the common maternal line and was crossed with the 200 other accessions through hand-pollination. Surface-sterilized seeds were maintained at 4 °C for 3 days in the dark and then sown on Murashige and Skoog plates containing 1% sucrose. All plants were grown under long-day conditions (16-h light at 22 °C and 8-h dark at 18 °C). For trait measurements, the plants were grown on Murashige and Skoog plates containing 1% sucrose for 7 days prior to transferring to soil for 7 days. The parental lines were grown alongside hybrids in the same pot with daily rotation. For RNA-seq, the plants were grown on Murashige and Skoog plates containing 1% sucrose for 14 days, and the first leaf from each plant was harvested for analysis.

Phenotyping

For leaf number measurements, only rosette leaves > 1 mm in length were included. For leaf area measurements, the area of the first leaf of each plant was measured using ImageJ software (http://rsb.info.nih.gov/ij/). The rosette diameter of each plant was manually measured. The shoot fresh weight of six plants was measured, and the average shoot fresh weight of each plant was calculated. Approximately 36 plants were measured for each genotype. Middle-parent heterosis (MPH) was calculated using the following equation: MPH = (F_1 – MP)/MP, where F_1 is the hybrid and MP is the mean of two parental lines; Better-parent heterosis (BPH) was calculated as BPH = (F_1 – BP)/BP, where F_1 is the hybrid and BP is the better-performing parental line.

Population genetic analyses

Approximately 8 million SNPs for 191 of the 200 accessions were downloaded from the 1001 Genomes Project website for *A. thaliana* (http://1001genomes.org/). SNP data were converted into a BEAGLE (1) (version 3.3.2, https://faculty.washington.edu/browning/beagle/b3.html) input file to impute genotype. Approximately 1.72 million SNPs were obtained after filtering out loci with triallelic SNPs and a minor allele frequency of < 0.05. The 1.72 million SNPs were pruned to 722,000 SNPs using PLINK1.07 software (2) (http://pngu.mgh.harvard.edu/~purcell/plink/) with parameters of window size in SNPs: 50, the number of SNPs to shift the window at each step: 5, the VIF threshold: 2, and these SNPs were then used to analyze the parental genetic distance and estimate Z (F_{ST}). Pairwise analysis of the parental genetic distance between Col-0 and each of the 191 paternal accessions was separately calculated using PLINK's IBD (identity by descent) analysis with the default parameters. For Z (F_{ST}) estimation, 191 accessions were divided into two groups according to biomass BPH or leaf area BPH: high-BPH group (the top third) and low-BPH group (the bottom third). F_{ST} was computed as previously described (3), using a 10-kb window and 5-kb sliding regions between the high-BPH group and the low-BPH group. The 10-kb window included the regions 5 kb upstream and downstream of each of the 750 associated SNPs. For the control, the genome of *A. thaliana* was randomly divided into 10-kb genomic regions, and F_{ST} was computed between the high- and low-BPH groups. Subsequently, F_{ST} was normalized using the Z score to obtain Z (F_{ST}).

RNA-seq and data analysis

RNA-seq was performed as previously described (4). Briefly, total RNA was extracted from the first leaf of Col-0, Per-1, Aa-0, Ak-1, Col-0×Per-1, Col-0×Aa-0, and Col-0×Ak-1 plants using an RNeasy Plant Mini Kit (Qiagen) at 14 days after sowing. The mRNA sequencing libraries were constructed, and sequencing was performed using the Illumina HiSeq 2500 platform. Two independent biological replicates were performed. RNA-seq reads of Col-0 were mapped to *Arabidopsis* TAIR 10 using TopHat software (5) (version 2.0.8b, http://ccb.jhu.edu/software/tophat/index.shtml) with parameters set to a minimum intron length of 20, maximum intron length of 6,000, and a maximum of two mismatches. To decrease the mapping bias for Per-1, Aa-0 and Ak-1 RNA-seq reads, we generated Per-1, Aa-0 and Ak-1 reference genomes using SNP data from the 1001 Genomes Project website for *A. thaliana* according to previously described methods (4). Then, RNA-seq reads of Per-1, Aa-0 and Ak-1 were mapped to Per-1, Aa-0 and Ak-1 reference genomes, respectively.

To analyze RNA-seq data from Col-0×Per-1, Col-0×Aa-0, and Col-0×Ak-1 plants, we aligned sequencing reads to both the Col-0 reference genome and updated Per-1, Aa-0 and Ak-1 reference genomes. Only unique mapped reads were extracted for the following analysis. The number of fragments per kilobase of exon model per million mapped reads (FPKM) for each gene, which represents the expression level of each gene, calculated using Cufflinks (6) (version 2.0.2, http://cole-trapnell-lab.github.io/cufflinks/). Using methods from a previous study (7), a threshold value of 1 was used for FPKM to identify expressed genes. Differentially expressed genes among two parents and the hybrid were identified using the generalized linear model (GLM) method in edgeR package (8) (version 3.12.0, http://www.bioconductor.org/packages/release/bioc/html/edgeR.html) with the FDR set at 0.05. For genes differentially expressed between two parents, when the expression level of a gene in the hybrid was significantly different from that in the low parent but not different from that in the high parent, then that gene was classified as ‘high-parent expression’, and if the expression level in the hybrid was significantly different from that in the high parent but not from that in the low parent, then that gene was classified as ‘low-parent expression’. GO analysis was performed using agriGO software (9).

All original data sets have been deposited in the Gene Expression Omnibus database under accession number GSE85759, GSE100595.

Quantitative RT-PCR

Total RNA was extracted as described above. A total of 3 μg of RNA was used as a template to synthesize cDNA using a RevertAid First Strand cDNA Synthesis Kit (Thermo). Quantitative RT-PCR was performed using a 7500 Fast Real-Time PCR Amplifier (Applied Biosystems) and SYBR Premix Ex Taq Kits (Takara). The expression of *UBC* was used as an internal control (10). At least two biological repeats for each experiment were performed. Quantitative RT-PCR reactions were performed with three technical replicates for each sample and the representative results are shown. The primers used for quantitative RT-PCR are listed in Table S3.

1. Browning BL, Browning SR (2009) A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet* 84(2):210–223.
2. Purcell S*, et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81(3):559–575.
3. Karlsson EK*, et al.* (2007) Efficient mapping of mendelian traits in dogs through genome-wide association. *Nat Genet* 39(11):1321–1328.
4. Yang M*, et al.* (2016) Natural variation of H3K27me3 modification in two *Arabidopsis* accessions and their hybrid. *J Integr Plant Biol* 58(5):466–474.
5. Kim D*, et al.* (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* 14(4):R36.
6. Trapnell C*, et al.* (2013) Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol* 31(1):46–53.
7. Sun X, Yang Q, Deng Z, Ye X (2014) Digital inventory of *Arabidopsis* transcripts revealed by 61 RNA sequencing samples. *Plant Physiol* 166(2):869–878.
8. Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26(1):139–140.
9. Du Z, Zhou X, Ling Y, Zhang Z, Su Z (2010) agriGO: a GO analysis toolkit for the agricultural community. *Nucleic Acids Res* 38(suppl_2):W64–70.
10. Czechowski T, Stitt M, Altmann T, Udvardi MK, Scheible WR (2005) Genome-wide identification and testing of superior reference genes for transcript normalization in *Arabidopsis*. *Plant Physiol* 139(1):5–17.

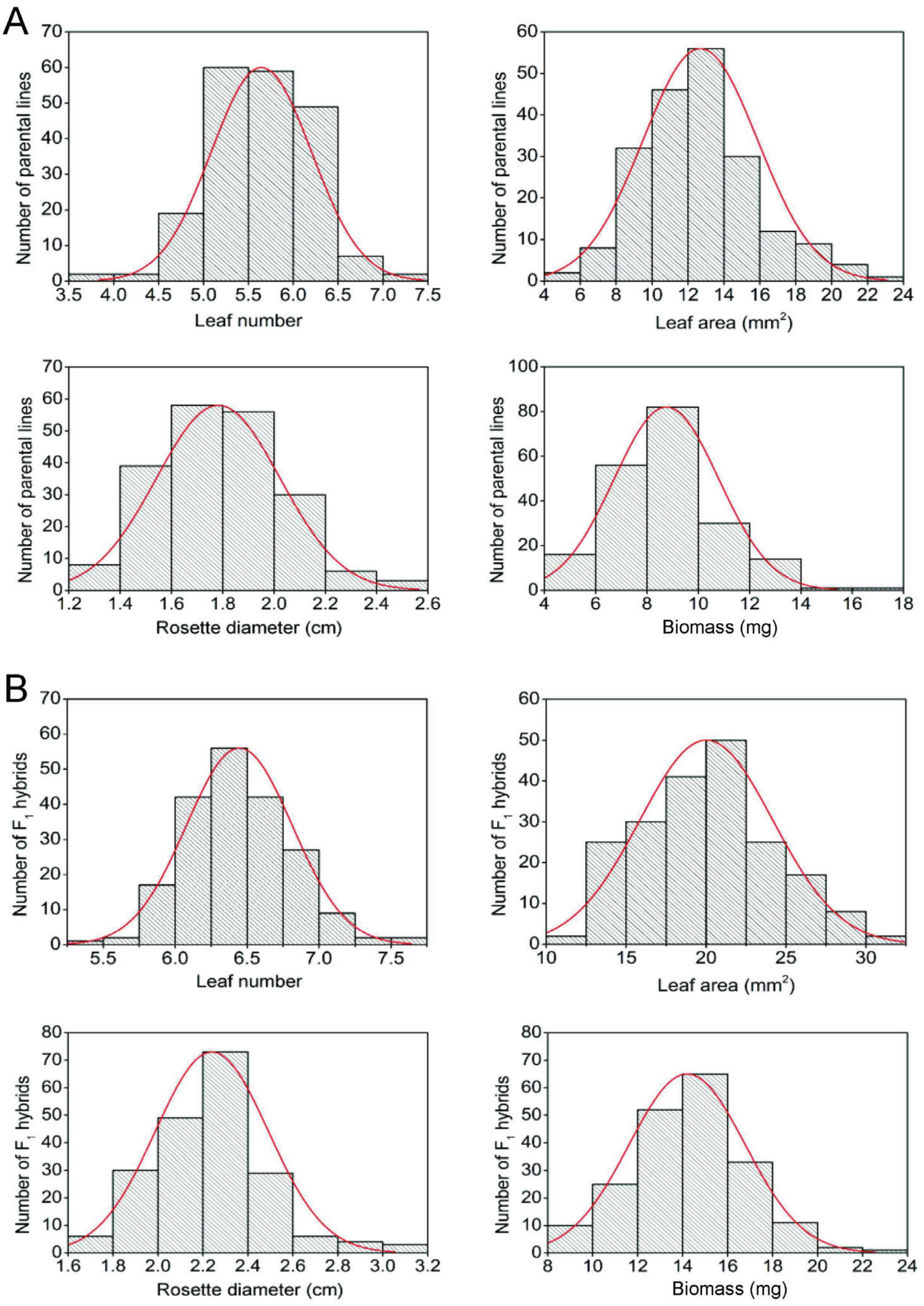


Fig. S1. Analysis of the four traits among 201 *A. thaliana* accessions (*A*) and among 200 *A. thaliana* F$_1$ hybrids (*B*) at 14 days after sowing (DAS).

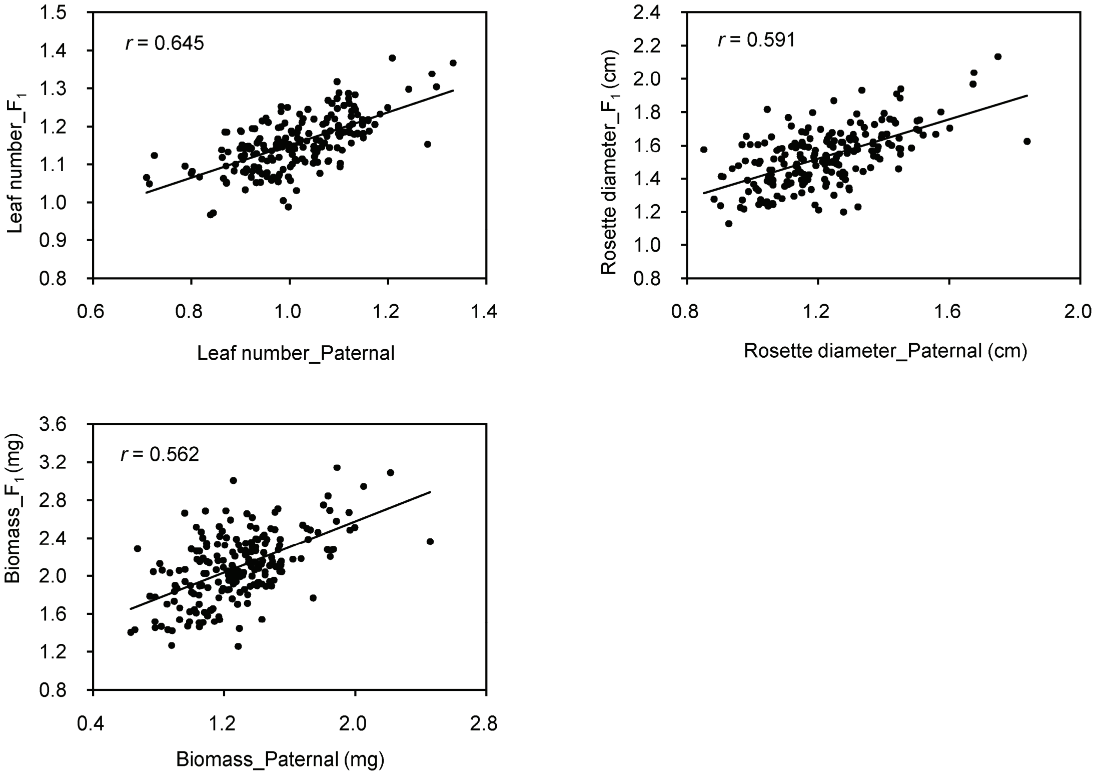


Fig. S2. Correlation of each trait between paternal accessions and hybrids.

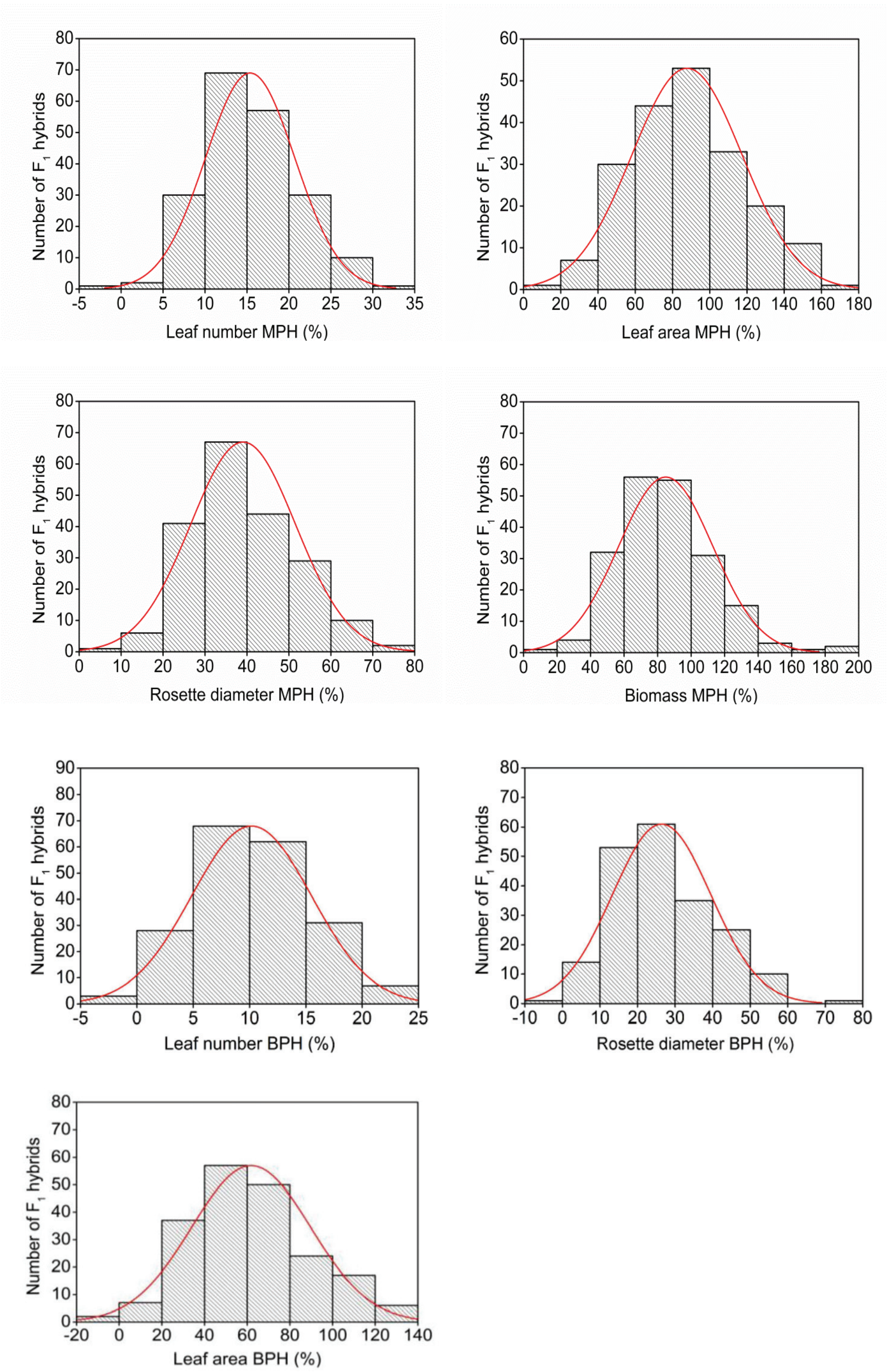


Fig. S3. MPH (middle-parent heterosis) and BPH (better-parent heterosis) in 200 *Arabidopsis* F_1 hybrids at 14 DAS.

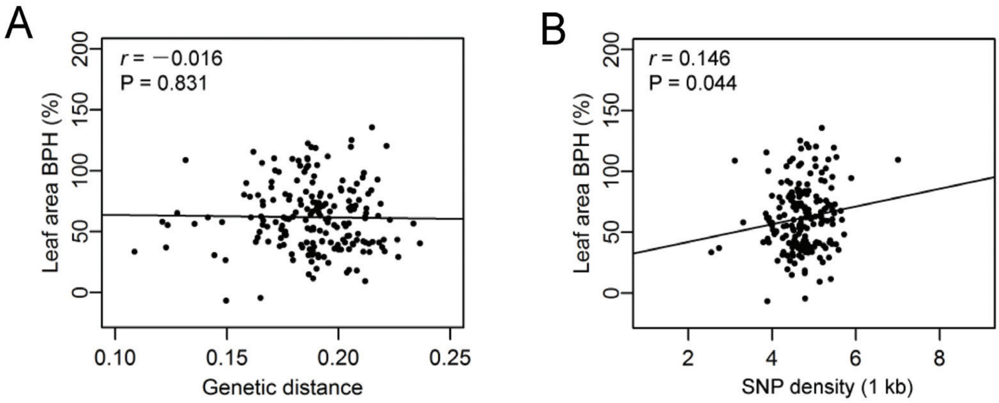


Fig. S4. Correlations between leaf area heterosis and parental genetic distance (*A*) or genome heterozygosity in *Arabidopsis* F_1 hybrids (*B*). The parental genetic distance was calculated between Col-0 and paternal accessions using the PLINK's IBD analysis. Genome heterozygosity in F_1 hybrids was represented as the SNP density per kilobase between both parents.

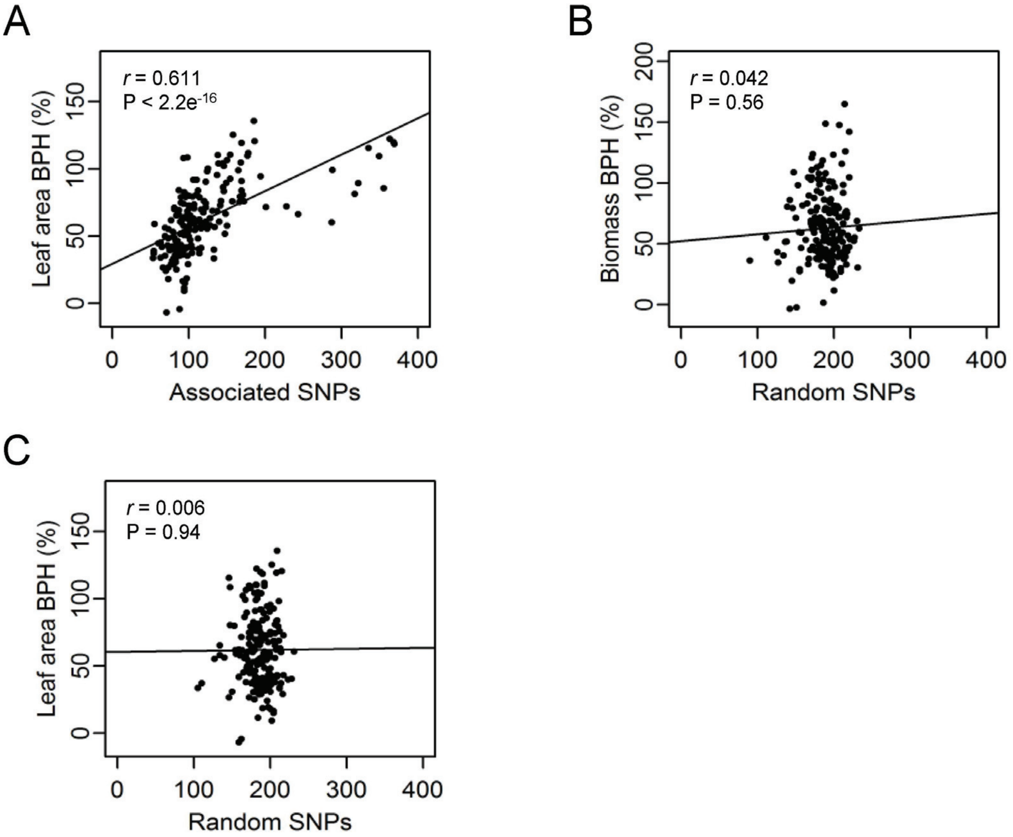


Fig. S5. Analysis of associated GWAS SNPs for heterosis in *Arabidopsis*. (*A*) Correlation of BPH for leaf area and the number of 750 associated SNPs accumulated in paternal accessions. Associated SNPs were identified using the Benjamini-Hochberg test with a threshold of 0.2. (*B*) Correlation of BPH for biomass and the number of 750 randomly selected SNPs accumulated in paternal accessions. (*C*) Correlation of BPH for leaf area and the 750 randomly selected SNPs accumulated in paternal accessions.

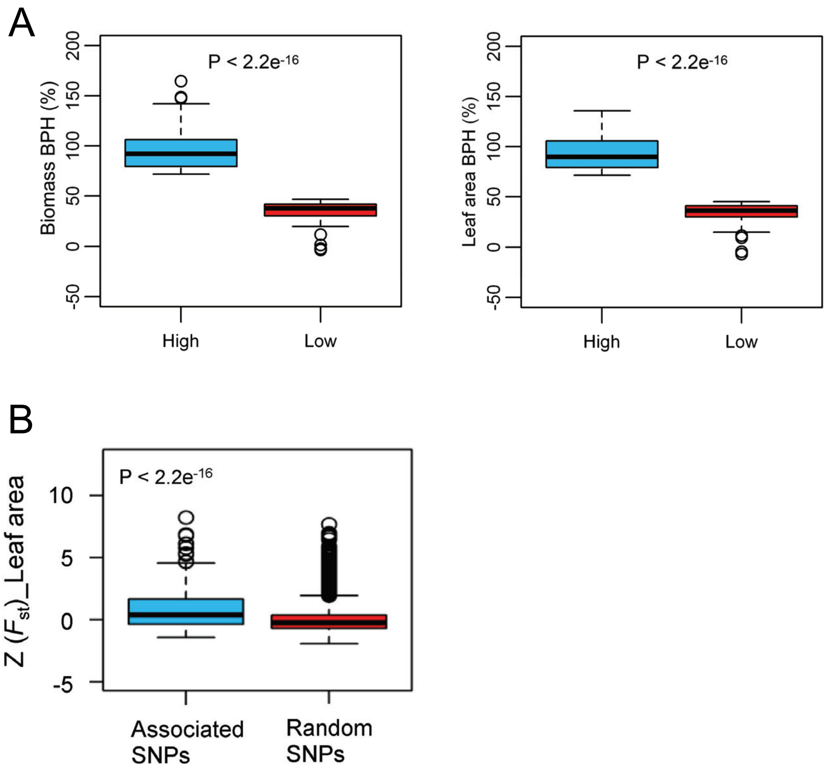


Fig. S6. The 10-kb genomic regions around 750 GWAS associated SNPs exhibited extensive sequence variation between the high-BPH and low-BPH groups. (*A*) Significant differences between the high- and low-BPH groups. *P*-values were obtained using Wilcoxon's rank-sum test. (*B*) Z(F_{ST}) value of the 10-kb genomic regions surrounding 750 associated SNPs between the high- and low-BPH groups for leaf area is significantly higher than that of the control. *P*-value was obtained using Wilcoxon's rank-sum test.

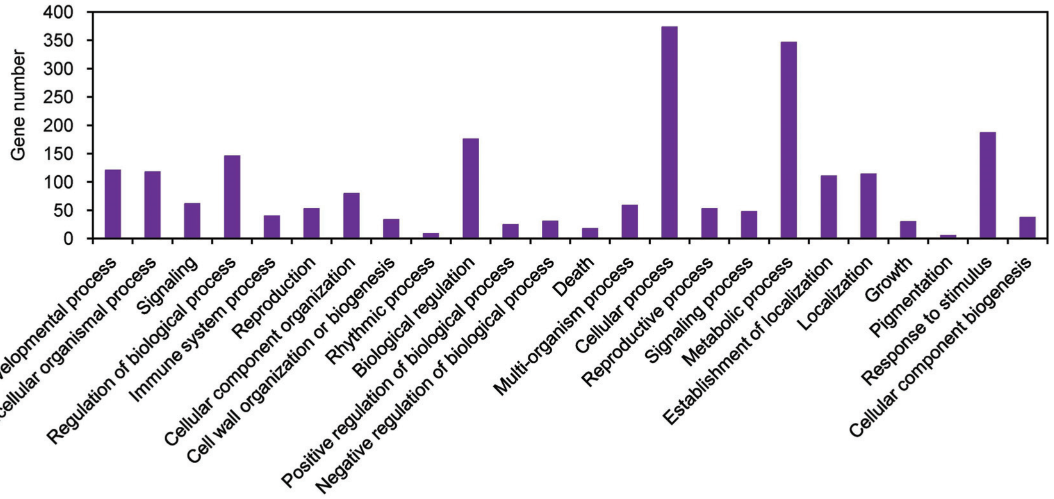


Fig. S7. GO functional analysis of candidate genes for biomass heterosis.

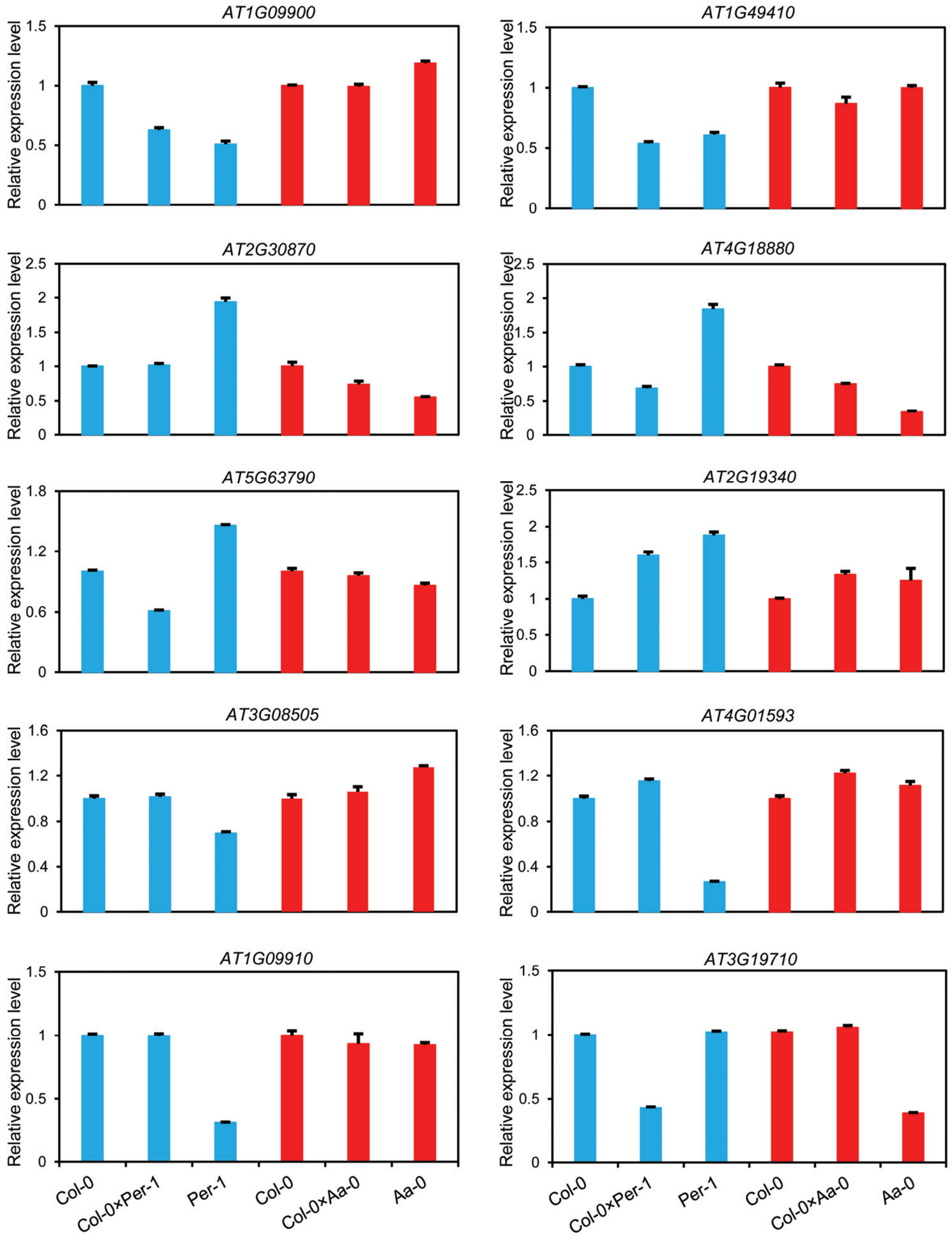


Fig. S8. Relative expression levels of candidate genes for biomass heterosis in Col-0×Per-1, Col-0×Aa-0 and their parents. *AT1G09900, AT1G49410*, *AT2G30870*, *AT4G18880* and *AT5G63790* exhibited low-parent or below low-parent expression in Col-0×Per-1 and mid-parent expression in Col-0×Aa-0. *AT1G09910, AT2G19340*, *AT3G08505* and *AT4G01593* exhibited high-parent expression in Col-0×Per-1 and mid-parent expression in Col-0×Aa-0. *AT3G19710* exhibited below low-parent expression in Col-0×Per-1 and high-parent expression in Col-0×Aa-0.

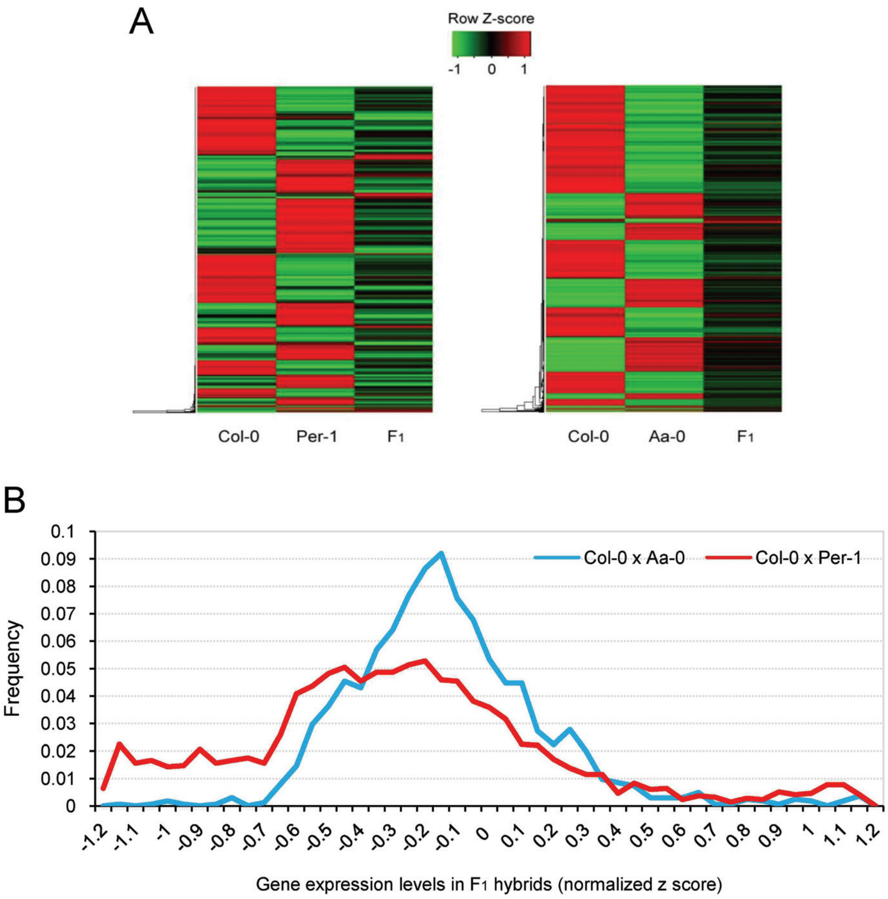
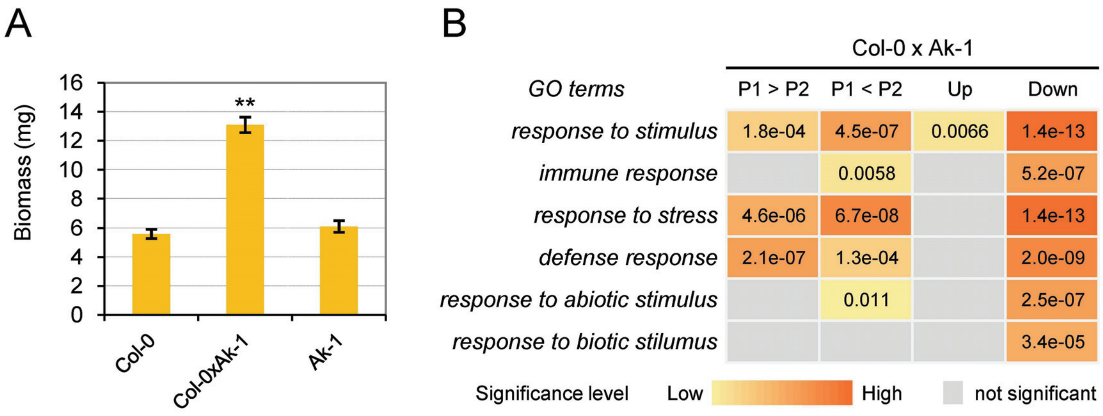


Fig. S9. Variation patterns of gene expression in Col-0×Per-1 and Col-0×Aa-0 hybrids. (*A*) Differentially expressed genes in two hybrid combinations. Gene expression data were scaled in row, and heatmaps were generated using Heatmap.2 in R. (*B*) The frequency of normalized Z score gene expression in F_1 hybrids. The lower the Z score, the greater the likelihood that F_1 hybrids will exhibit lower gene expression levels than the low parent.



	Col-0 x Ak-1			
GO terms	P1 > P2	P1 < P2	Up	Down
response to stimulus	1.8e-04	4.5e-07	0.0066	1.4e-13
immune response		0.0058		5.2e-07
response to stress	4.6e-06	6.7e-08		1.4e-13
defense response	2.1e-07	1.3e-04		2.0e-09
response to abiotic stimulus		0.011		2.5e-07
response to biotic stilumus				3.4e-05

Fig. S10. Transcriptomic analysis of hybrid Col-0×Ak-1 with high biomass BPH. (*A*) Col-0×Ak-1 exhibits high BPH for biomass. The data are presented as the mean ± SD; n > 30. ** $p < 0.01$ between hybrid individuals and parents (Student's *t*-test). (*B*) GO analysis of genes with expression variations in Col-0×Ak-1.

Table S1. Correlation of traits in F_1 hybrids

r	Leaf number	Leaf area	Rosette diameter	Biomass
Leaf number	1			
Leaf area	0.051	1		
Rosette diameter	0.06	0.847	1	
Biomass	0.435	0.822	0.769	1

Table S2. Correlation of BPH in F_1 hybrids

r	Leaf number	Leaf area	Rosette diameter	Biomass
Leaf number	1			
Leaf area	0.48	1		
Rosette diameter	0.494	0.86	1	
Biomass	0.598	0.886	0.859	1

Table S3. Primers for quantitative RT-PCR used in this study

Gene	Forward primer (5'-3')	Reverse primer (5'-3')
AT1G09900	GAGCGATTGACATATTGGAGAAGA	TGTTAGCATAGTGTTGTAGGTGAC
AT1G09910	CAGCCTCATCCGCAATAGC	GTGGTAGATTACTTATCCGTGACA
AT1G49410	AGCAGCACGGATAATGACAAC	CCCAGGAATGTTCATGCGAAA
AT2G19340	ATCGGAGAAGGAAGCGTGAA	CCAGTGTATCAGACGGTTCAG
AT2G30870	GAGAGGACAAGTAGAGCAATGG	CACCAGCCAAGTATTCGTTCT
AT3G08505	CACCGTCATTATTATCGCCTTGT	GCTGCTTCCAATCTTCCTCTTAG
AT3G19710	GAACCAAGAATCGGACCACTC	GCTGACAGACTCTATATGCCTTAT
AT4G01593	GCATACTCTGGTTTAATCACTCAC	GTTACAACAATCTCACGGCATT
AT4G18880	TGCTACTGGAGGAGGAGGAG	ATCTTGTATCGGATTCTTGTGAGAG
AT5G63790	AATCGGAGGACTGGTACTGAA	GTCGGACTCAAGATGCTCAAG
UBC	CTGCGACTCAGGGAATCTTCTAA	TTGTGCCATTGAATTGAACCC

Other Supporting Information Files

Dataset S1 (XLS)

Dataset S2 (XLS)