Supporting Online Material for

Common Sequence Polymorphisms Shaping Genetic Diversity in *Arabidopsis thaliana*

Richard M. Clark, Gabriele Schweikert, Christopher Toomajian, Stephan Ossowski, Georg Zeller, Paul Shinn, Norman Warthmann, Tina T. Hu, Glenn Fu, David A. Hinds, Huaming Chen, Kelly A. Frazer, Daniel H. Huson, Bernhard Schölkopf, Magnus Nordborg, Gunnar Rätsch, Joseph R. Ecker, Detlef Weigel*

*To whom correspondence should be addressed. E-mail: weigel@weigelworld.org

Published 20 July, *Science* **317**, 338 (2007)
DOI: 10.1126/science.1138632

This PDF file includes:

SUPPORTING ONLINE MATERIAL

TABLE OF CONTENTS

MATERIALS AND METHODS

1. ARRAY DESIGN

The entire 119,186,497 bp *A. thaliana* genome (*S1*) from accession Col-0 was used as the *reference sequence* for array design without repeat masking. In total, 118,991,806 bp in the reference genome assembly with unambiguous base calls (i.e., A, G, C or T) were included within 1-bp tiling paths suitable for polymorphism discovery (Fig. S1). These self-designed arrays were synthesized by Affymetrix (Santa Clara, CA, USA) with photolithography in conjunction with chemical coupling to direct the synthesis of the 25-mer oligonucleotides (*S2-S4*). The features were distributed over five large microarray (wafer) designs (see also Section 16).

2. SAMPLE PREPARATION AND HYBRIDIZATION

Isolation of genomic DNA

For each of 20 *Arabidopsis thaliana* accessions (Table S1), genomic DNA was prepared from ~8 g of leaf tissue collected from 2-6 week old plants grown at 23˚C under long days (16 hours light) with a modified version of a Qiagen (Valencia, CA, USA) user defined protocol. Either freshly collected leaves or leaves stored at -80˚C were ground in liquid N_2 to a fine powder with a mortar and pestle, and 4 ml of powder (~2 g) was placed in 50 ml tubes containing 27 ml digestion buffer [20 mM ETDA, 10 mM Tris-Cl, pH 7.9, 1% Triton X-100, 500 mM guanidine-HCl, 200 mM NaCl, and 4 g/L Driselase (Sigma-Aldrich, St. Louis, MO, USA, D9515)]. Samples were next incubated at 39˚C for 2 hours, and gently inverted every 30 minutes. 20 µl DNAse-free RNAse A (20 mg/ml, Fermentas Life Sciences, Burlington, Ontario, Canada, EN0531) was then added to each tube, and samples were incubated for an additional 30 minutes at 37˚C, followed by addition of 500 µl Proteinase K (50 U/ml, Roche, Basel, Switzerland, Cat. No. 3115844), and incubated at 55˚C for 2 hours with gentle inversion every 30 minutes. Samples were spun at 11,900 × g to pellet cellular debris, and supernatants for a given accession were combined and filtered through two layers of Miracloth to remove residual particulate matter. The resulting solution was applied to Genomic-tip 100/G columns (Qiagen, Cat. No. 10243) equilibrated with 4 ml of Qiagen buffer QBT (3-4 columns were used per accession). Columns were then washed three times with 7.5 ml of Qiagen buffer QC that had been preheated to 55˚C, and DNA was eluted with 7.0 ml of Qiagen buffer QF preheated 55˚C. DNA was precipitated by the addition of 5 ml room temperature isopropanol, and pelleted at 5,000 × g for 40 minutes. Pellets were washed with 5 ml of 70% ethanol, and spun at 5,000 × g for 40 minutes. The resulting pellet was air dried, and genomic DNA was resuspended overnight at 4˚C in 150-200 µl of sterile water.

Whole-genome amplification and labeling of DNA for hybridization

To generate sufficient DNA for hybridization, each DNA sample was whole-genome amplified with the Repli-g kit from Qiagen. This whole-genome amplification was carried out as recommended by the manufacturer in a scaled up to a reaction volume of 25 ml created by combining the contents of 5 kits for each sample. The whole-genome amplified DNA samples were precipitated with the addition of 0.1 volume of 3M sodium acetate (pH 5.5) and 0.7X

isopropanol, transferred to 15-ml tubes, washed twice with 80% ethanol and dried at 70˚C for ~15 minutes. Samples were resuspended in 5 ml of 10 mM Tris (pH 8.0) and incubated at 60˚C for 15 minutes with periodic vortexing. To remove residual precipitate, the samples were spun at ~11,000 × g for 5 minutes at room temperature, and the supernatant was transferred to a 15 ml tube. Any remaining precipitate was removed by spinning aliquots of the supernatant at 20,800 × g for 4 minutes in 1.5 ml tubes, before recombining the aliquots back into a 15 ml tube. DNA concentration was measured with a spectrophotometer with 1:150 dilutions in sterile water.

Each amplified DNA sample (2.7 – 2.8 μg/μl) was fragmented for 8 minutes at 37˚C in a total of 6430 μl of the following reaction mixture: 1X One-Phor-All Buffer PLUS (Amersham, Piscataway, NJ) and 0.016 mM DNase I (pH 8.0) (Invitrogen, Carlsbad, CA). DNase I was heat-inactivated at 99˚C for 5 minutes. This protocol resulted in a peak fragment size of 100 bp. The fragmented samples were labeled for 90 minutes at 37˚C in a total of 8,410 μl in the following reaction mixture: 0.16 mM biotin-16-[ddUTP + dUTP] (Perkin Elmer, Boston, MA), 21.4 U/μl rTDT (Roche Applied Science, Indianapolis, IN) and 0.21X One-Phor-All Buffer PLUS. rTdT was heat-inactivated by incubation for 10 minutes at 99˚C.

Array hybridization

A total of 21 ml of hybridization mix, containing the following reagents, was prepared: 8410 μl labeled target DNA, 2.92 M tetramethylammonium chloride, 0.01 M Tris pH 7.8, 0.01% Triton X-100, 0.05 nM control oligo b-948 (Proligo, Boulder, CO), 0.1 μg/μl herring sperm DNA (Promega, Madison, WI), and 0.5 mg/ml acetylated bovine serum albumin (BSA). For each sample the first array was hybridized to 14 ml of the hybridization mix at 50˚C for 18 hours. The hybridization mix was then removed and reused for hybridization to the second array. It was then supplemented with the remaining 6 ml and reused consecutively for the remaining three arrays. Hybridizations were performed in an Affymetrix oven with a rotation speed of 10 revolutions per minute. After hybridization, arrays were washed at high-stringency in 0.2 volumes of SSPE, 0.01% TX-100 for 60 minutes at 37˚C.

Hybridized DNA probe was detected by incubation with the following series of reagents: 5 ng/μl streptavidin (Invitrogen) for 20 minutes, 2.5 ng/μl biotinylated anti-streptavidin (Vector Labs, Burlingame, CA) for 20 minutes, 1 ng/μl streptavidin-Cy-chrome (Pharmingen, San Diego, CA) for 20 minutes, 2.5 ng/μl biotinylated anti-streptavidin for 10 minutes, and 1 ng/μl streptavidin-Cy-chrome for 10 minutes at room temperature. A final high-stringency wash was performed in 0.2X SSPE, 0.01% Triton X-100 at 37˚C for 1 hour if needed. Arrays were scanned with custom-built confocal scanners.

3. EXPERIMENTAL INPUTS FOR PREDICTION ALGORITHMS

Fluorescence intensity data from the array scans were first processed to determine an average intensity I for each feature on the array. This yields 8 data points per sequence position, one each for A, C, G, and T on each of the forward and reverse strands. For each position a "raw base call", denoted B, was defined as the base corresponding to the nucleotide probe that showed the highest intensity among the four probes for a given strand and accession. Quality scores, denoted by Q, were computed for each position in each accession for both strands with an algorithm similar to *Phred* (*S5*) that considers the ratio of the highest and second highest intensities and the conformance of surrounding base calls with the reference sequence. The scoring algorithm

derives a decision tree for estimating error rates for individual raw base calls on the basis of the input metrics. Since these trees are made from a limited number of nodes, a limited set of discrete scores is possible. Similar to dideoxy sequencing quality scores, the reported scores represent estimated base 10 log error rates (e.g. Q=20 corresponds to an error rate of 0.01, Q=30 to an error rate of 0.001, etc.). The quality scores were calibrated with scans of the Col-0 accession. Due to experimental variation between hybridization experiments, the quality scores for an individual scan may not be perfectly calibrated, and may systematically underestimate or overestimate error rates. While quality scores were not used directly in the SNP calling algorithms we present, they were employed for prediction of polymorphic regions and reference base calls (see Sections 8 and 9).

4. WHOLE GENOME ANNOTATION OF REPETITIVE PROBE SETS

Cross-hybridization of repetitive sequences confounds polymorphism detection from oligonucleotide arrays, and can either (i) mask legitimate polymorphisms or (ii) introduce anomalous intensity readings for nonpolymorphic regions that lead to spurious polymorphic predictions. For each tiled position, we therefore determined whether probes match with high sequence complementary to additional genomic locations. We subsequently used this information in the algorithms described below or for *ad hoc* curation of predictions.

Exact, short, and inexact 25-mers matches

We distinguish 3 classes of matches between repetitive 25-mer probes, each of which is allowed a mismatch at the central (13^{th}) position that varies as part of the array design (Fig. S2). First, *exact 25-mer matches* correspond to probes that are completely complementary to at least two genomic locations (on either genomic strand) for positions 1-12 and 14-25. Second, because mismatches at the ends of probes have comparatively little effect on hybridization strength (*S6*), we identified *short 25-mer matches* according to the same rules except that mismatches were allowed on any or all of the 2 bp on either end of 25-mer probes. Finally*, inexact 25-mer matches* correspond to probes that have multiple complementary counterparts in the genome with one mismatch at positions 1-12 or 14-25. For inexact matches, the potential for stable duplex formation (and for cross-hybridization on arrays) is more difficult to predict, and is expected to vary depending on sequence properties and mismatch location within the probe (*S6*).

The entire Col-0 reference genome sequence was used for 25-mer annotation, as were the chloroplast and mitochondrial genomes that were a contaminant in genomic DNA preparations used for hybridization to arrays. Briefly, we generated a list that contained 25-mers with a 1-bp tile of the forward and reverse strands of the entire nuclear and organellar genomes. Each 25-mer was identified by its genomic location (i.e. the location of its center position). In a second step this list was sorted according to the nucleotide sequence, and 25-mers occurring more than once were extracted from the sorted list in a linear traversal.

The sorting algorithm was then modified to handle mismatches. We used a recursive, position-wise partitioning method that begins by partitioning the tiling list according to the nucleotide at position 1 of each 25-mer. This partition is then recursively subdivided according to subsequent positions. Mismatches at the central 25-mer position are tolerated by skipping the 13th partitioning step. Partitions created when sorting on position 12 are therefore subdivided according to the nucleotides at position 14. The generalization of the sorting method to short 25-

mer matches is straightforward: in addition to position 13, positions 1, 2 and 24, 25 are skipped.

The class of inexact 25-mer matches can be seen as a (disjoint) union of 20 subclasses each containing matches with two fixed mismatch positions i and 13, where subclass index $i \in \{3,4,...,12, 14,...,22,23\}$. Each subclass of inexact 25-mer matches can be easily computed with our approach by skipping a pair of fixed positions (i,13). After independently running the whole sorting and parsing procedure 20 times, we took the union of the resulting matches to obtain the whole class of inexact 25-mer matches.

As 25-mers had been tagged with genome locations, mapping final partition blocks back to the genome was straightforward. Counts of positions with exact, short, and inexact 25-mer matches are given in Table S2. We also identified a subset of positions with matches elsewhere in the genome for which the counts of the nucleotide at the central position exceeded the perfect match central position. These *dominating 25-mer positions* are especially likely to lead to false SNP predictions. Information for these dominating positions was used by the learning algorithms for SNP prediction as described in Section 6.

5. A MODEL BASED METHOD (MB) FOR SNP IDENTIFICATION

SNP prediction with model based method

We used the same pattern recognition algorithms for analysis of the *A. thaliana* resequencing data that had previously been developed for array-based resequencing and SNP discovery in the human genome (*S3, S4*).

Intensity measurements (I), as well as the raw base calls (B), were employed as inputs to the MB algorithm. We also determined the local "conformance" of the array data, as the fraction of base calls that matched the reference sequence within a sliding window. For a position where the direct call matched the reference base, this window consisted of bases at positions −10 to +10. In the immediate vicinity of an alternate base call, hybridization intensities are reduced due to the presence of a one-base mismatch base between the target and probe DNA. To avoid the reduced-intensity interval in these cases, we altered the window to span bases −20 to −10, and +10 to +20. A strict base call was made for a sequence position when the ratio of the brightest to next-brightest feature was greater than a threshold of 1.3, and the conformance around that position was at least 0.80. For alternate base calls that did not match the reference sequence, we also required that there were no brighter alternate calls meeting these criteria within positions −5 to +5. For polymorphism detection we used these strict-called sequences to create a consensus sequence of calls that were confirmed on both strands. Again, alternate consensus calls were excluded if there was a brighter (average intensity over both strands) alternate consensus call within positions −5 to +5. Putative polymorphic sites were also required to pass a final "footprint test". In this test, normalized intensities for probes matching the reference sequence across positions −5 to +5 were separately averaged for scans that resulting in reference base calls and alternate base calls. The normalization step adjusted for systematic differences in brightness between scans. A SNP was rejected if the ratio of mean normalized intensity around reference calls to mean normalized intensity around alternate calls was less than 1.5. The footprint test required a cumulative analysis of a complete set of arrays of the same design. We required at least one consensus reference call and one alternate call to define a polymorphism; positions with no reference calls were rejected. Once a site was determined to be polymorphic in at least one accession, we relaxed the base calling criteria and accepted strict calls on just one strand if the

other strand was found to be ambiguous (i.e., did not pass either the intensity ratio or conformance requirements). Predictions at positions with exact and short 25-mer matches, where the potential for cross-hybridization was high, were subsequently removed.

Estimating performance for MB SNP calls

The 19 non-Col-0 accessions hybridized to arrays are a subset of accessions sampled by PCR amplification and dideoxy sequencing of ~500-600 bp regions throughout the *A. thaliana* genome as part of the NSF-funded Arabidopsis 2010 project (*S7*). In addition to previously published sequences (*S7*), unpublished data that are freely available for download were used (see Section 16).

We used sequence information from 1,213 fragments (available as of July 26, 2005) to assess SNP prediction accuracy and recall for the MB method as well as for additional methods described in the following sections. While the Van-0 accession was included in the 2010 dataset ("2010"), the presence of extensive heterozygous SNP calls relative to the other accessions precluded accurate error assessment. A seed stock of Van-0 ascertained by genome-wide scans for several hundred SNPs to be homozygous throughout the genome was kindly provided by J. Borevitz (Univ. of Chicago), and used in this study.

Absolute numbers for MB SNP predictions per accession, with FDRs and recall established with 2010, are provided in Table S3 (see column "MB"). Recovery by the MB method was not strongly influenced by allele frequency (Fig. S3), and for the Col-0 reference, we predicted 470 SNPs genome-wide. These may be either false positive predictions from the array data, or incorrect base calls for the reference sequence. We also assessed calling accuracy for MB predictions at sites of inexact 25-mer matches that we did not exclude in making predictions with the MB method. While the number of such test examples in 2010 is low (249 predictions at these sites across all accessions), the resulting FDR of ~6.7% is about 3.4X higher than for all MB predictions. Of the 449,468 positions included in the MB SNP dataset, 4.2% have inexact 25-mer matches. We lack data from 2010 to assess the rate at which the MB method generates predictions in large deleted regions. However, for a set of validated deleted bases in the target accessions (Section 11), most of which were in deletions greater than ~300 bp, we observed 11 predictions by the MB method at a total of 132,407 deleted bases (1 false MB call per every ~12 kb in deleted regions).

Finally, we also assessed the FDR for reference base calls at positions predicted by the MB method to harbor a substitution in at least one other accession. For 41,655 reference calls in the MB dataset for which information was available from 2010, the rate of false assignment for reference base calls was 0.031%.

6. A MACHINE LEARNING (ML) METHOD FOR SNP IDENTIFICATION

To complement and extend the set of SNP predictions from the MB approach (Section 5), we implemented a novel method to predict SNPs from array data. This method uses machine learning (ML) methodology and features Support Vector Machines (SVMs). Machine learning methods rely on known datasets both for training and error evaluation. Such a known dataset, the 2010 dataset, was available. The absence of reliable data from the Van-0 accession from the 2010 dataset precluded the use of ML methods for Van-0. Finally, because information from the Col-0 reference accession was used for training SVMs, the ML algorithms could not be applied

to hybridization data from the Col-0 accession itself (e.g., to identify potential sequence errors in the published reference sequence).

In a first step SVMs were trained on a per-accession basis with array data from a given accession and the Col-0 accession, as well as the reference sequence (layer 1 SVMs). In a second step, we exploited information across all accessions in training a second set of SVMs (layer 2 SVMs), which were used to make final predictions. Again, training was performed on a per accession basis. For both layer 1 and 2 SVMs, we performed five subtasks: (i) position filtering, (ii) input generation, (iii) model selection and training, (iv) prediction, and (v) transformation of output values. In the final step, we assigned confidence values to each prediction reflecting the likelihood of a true SNP prediction. A cross-validation procedure was employed to obtain unbiased confidence estimates (i.e., data points used for training or model selection were excluded in assessing prediction precision). The details of this method are described below, and an overview of the method is shown in Fig. S4.

Layer 1 SVMs

Filter for layer 1 SVMs

Prior to training layer 1 SVMs, we excluded positions which were either (i) likely to be non-polymorphic in a given accession, or (ii) were likely to correspond to positions with intrinsically poor probe-set properties. For SNP prediction in a given target accession t, we exclusively considered positions p_t satisfying the following conditions. First, raw base calls $B^+_{Col}(p)$ and $B^-_{Col}(p)$ on the forward and reverse strand of the Col-0 accession had to correspond to each other and to the expected base call for the reference sequence $seq(p)$. Secondly, there had to be identical alternate raw base calls $B^+_t(p)$, $B^-_t(p)$ in the target accession t on both strands. Formally,

$$p_t = \{p \mid B^+_{Col}(p) = B^-_{Col}(p) = seq(p) \wedge B^+_t(p) = B^-_t(p) \neq seq(p)\}.$$

Finally, as positions corresponding to dominating 25-mer matches are likely to be particularly problematic for SNP prediction (see Section 4), these positions were rejected. After applying the filter, 99% of all positions were excluded as SNP candidates, including ~30% of positions with true SNPs (estimates based on the 2010 dataset; see Table S4). Thus, the ratio of positive examples (true SNPs) to all examples (any position) is reduced from ~1:230 to ~1:4. This provides a more balanced dataset and saves computational time for both training and prediction.

Input generation for SVMs

For each position p passing Filter 1 in a target accession t we generated an input vector $\boldsymbol{x}^{(1)}$ by concatenating measurements at this position and at neighboring positions ± 4 bp from p. This feature vector is defined as:

$$\boldsymbol{x}^{(1)} = [I_{max}, I_{sec}, Q_1, Q_2, k, M, seq, f, S].$$

It includes maximal intensities I_{max} and averages of the non-maximal intensities I_{sec} for every position in the 9 bp window, quotients Q_1 corresponding to the ratios of the maximum intensities at p and its neighboring positions, quotients Q_2 corresponding to the maximum intensities of the

target and the Col-0 accession, occurrences of probes *k* within the 9 bp neighborhood with matches at multiple genomic locations (see Section 4), mismatches *M* between raw base calls and the reference sequence within the 9 bp neighborhood, the reference base *seq* at the considered position, frequencies *f* of each letter of the alphabet (A, C, G, T) within each probe and the sequence entropy *S* of the probe. A detailed description of all inputs is provided in Table S5.

After normalization of the input vectors on the training set (mean 0, standard deviation 1, per input dimension), the vectors were employed in SVMs (*S8, S9*). For the training data $(\mathbf{x}_i, y_i)$, we used the corresponding output labels for the given target accession *t* with $y \in \{-1, 1\}$, i.e. "no SNP" and "SNP", respectively. On the basis of *n* labeled examples we used SVMs to learn a discriminate function

$$F(\mathbf{x}) = \sum_{i=1}^{n} y_i \alpha_i k(\mathbf{x}, \mathbf{x}_i)$$

parameterized by α. It uses a so-called kernel function $k(\mathbf{x}_i, \mathbf{x}_j)$ computing the similarity of the two vectors $\mathbf{x}_i$ and $\mathbf{x}_j$. Here we used the standard radial basis function (RBF) kernel:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\left\|\mathbf{x}_i - \mathbf{x}_j\right\| / 2\sigma^2\right)$$

with hyper-parameter σ. The variables α are determined by solving the following SVM optimization problem (*S8, S9*):

$$\begin{aligned} \min \quad & \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) + C_+ \sum_{i:y_i=+1} \xi_i + C_- \sum_{i:y_i=-1} \xi_i \\ s.t. \quad & y_i \sum_{j=1}^{n} y_j \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) \geq 1 - \xi_i \\ w.r.t. \quad & \xi_i \geq 0, \alpha_i \geq 0 \qquad \text{where} \quad i = 1, \ldots, n. \end{aligned}$$

Here, the hyper-parameters C_+ and C_- determine the trade-off between margin maximization and error minimization as well as the trade-off between false positive and false negative predictions. The additional variables ξ_i are slack variables allowing for outliers in the training set. The kernel parameter σ was tuned during model selection along with the hyper-parameters C_+ and C_-. For fast and efficient training and prediction of SVMs we used the SHOGUN toolbox, developed by Sonnenburg and colleagues (*S10*).

Cross-validation and model selection

To perform the three tasks of (i) training, (ii) model selection, and (iii) evaluation of the generalization error, the labeled 2010 dataset was divided into three disjoint sets. The first set was used for training with *k* different models; the second set served for tuning of the model parameters, and the generalization error was computed on the third set. To minimize statistical errors during the evaluation we predicted each position in the labeled set with an SVM that had not seen the example during training or parameter tuning. The instances of the three sets were therefore permuted through 2010 in 5-fold cross validation (Fig. S5) (*S11*): the 2010 dataset was randomly split into five disjoint sets of equal size, and model selection and training were performed 5 times on sets $X_m = 2010 \setminus T_m$ (where "\" denotes the set difference), each time with a different set reserved as test set T_m, with $m=1,\ldots,5$.

For the model selection each training set X_m, which contained 80% of all labeled samples, was again subdivided into 5 disjoint sets. For each set X_m we trained 5 times for each model *k* on

subsets $X_{mn} = X_m \backslash T_{mn}$, each time leaving out one subset T_{mn}. The predictions on the omitted subset T_{mn} were then used to choose the best model. For that purpose we calculated the number of false positives, *FP*, as a function of the number of true positives, *TP*. The proportion of *FP* to *TP* can be assessed with respect to a given decision threshold on the output space (Fig. S6). As optimization criterion for the model selection we determined the area a_{mnk} between the computed curve *FP*=*FP*(*TP*) and a line representing 1 FP at 50 TP (Fig. S7). For each set X_m, the model k_m which maximizes the sum over the areas a_{mnk} of the five subsets T_{mn} with $n=1...5$, was considered optimal. With these criteria, we optimized over a range of acceptable, low FDRs suitable for biological studies.

As we used Gaussian RBF kernels, the parameters to be tuned included the width σ ($\sigma=[10^2, 10^{2.3}, 10^{2.7}, 10^3, 10^{3.3}, 10^{3.7}, 10^4]$) and the *C*-values ($C_+ =[10^{-0.1}, 10^{0.25}, 10^{0.6}, 10^{0.95}, 10^{1.3}]$, *and* $C_- = [0.2,0.4,0.6] \times C_+$). In total 105 models were tested. Having chosen the model k_m, the whole set X_m was trained with this model and the predictions were computed for the left out set T_m. At the end of this procedure there were 5 different SVMs for the accession *t,* trained each with a different model k_m. As we also used the subsets T_m for the calibration of the SVM output values (see below), we did not retrain on the whole labeled set.

Prediction

For each position p_t in 2010 that passed filter 1 in accession *t*, exactly one prediction $F^m{}_t(p_t)$ was computed with the single layer 1 SVM that had not seen the example p_t during training or parameter tuning. As the rest of the genome was not employed in training or tuning, any SVM trained on the corresponding accession could be used. Therefore, for each unlabeled site, one of the 5 layer 1 SVMs was randomly chosen.

Transformation of SVM output values into confidences

The predictions of the five layer 1 SVMs $F^m{}_t$ for each of the 18 accessions were based on different models and therefore were not directly comparable. To combine the outputs for use in subsequent analyses, we scaled the outputs relative to each other by assigning to each prediction a probability for being a true positive (i.e., a correctly called SNP). Both tasks can be resolved by estimating the conditional likelihood $P(y_t=1|F^m{}_t)$ of the true label y_t being positive for a given output value $F^m{}_t$ of the layer 1 SVM.

To do this, we applied a piecewise linear function which was determined on the corresponding validation set T_m. We used the 1/20 quantiles taken on the SVM output values as supporting points *x(l)* (Fig. S8). For each point *x*(*l*) the corresponding $\bar{y}$-value, which represents the probability of being a true positive, was computed as:

$$\bar{y}(l) = \frac{n_{TP}(l)}{n(l)},$$

where *n*(*l*) is the number of examples in 2010 with output values $x(l) \le F_t^m \le x(l+1)$, and $n_{TP}(l)$ is the sum of labeled SNPs in the same output range. We additionally defined a cumulative probability function $\bar{y}_c$, which is the mean probability for all positions with output values $F^m{}_t \ge x(l)$:

$$\bar{y}_c(l) = \frac{n_{c,TP}(l)}{n_c(l)},$$

where $n_c(l)$ and $n_{c,TP}(l)$ are similarly defined as $n(l)$ and $n_{TP}(l)$ with output values $F^m{}_t \geq x(l)$. We applied a technique to obtain smooth and monotonically increasing estimates (available on request).

For any output value $F^m{}_t$ the corresponding confidence c is then given by linear interpolations:

$$c = \begin{cases} y(1), \text{ for } F_t^m \leq x(1) \\ \dfrac{y(l+1)\cdot(F_t^m - x(l)) + y(l)\cdot(x(l+1) - F_t^m)}{x(l+1) - x(l)}, \text{ for } x(l) \leq F_t^m \leq x(l+1) \\ y(20), \text{ for } F_t^m \geq x(20) \end{cases}$$

and similarly for the cumulative confidence C with corresponding $\bar{y}_c$. Each predicted output value was transformed with the piecewise linear function corresponding to the layer 1 SVM used.

Layer 2 SVMs

Filter for layer 2 SVMs

For further analysis in layer 2 SVMs, we excluded all positions where the transformed layer 1 SVM outputs c_a for all 18 accessions a scored below an appropriately chosen threshold K_a. At positions that were likely to have a SNP in at least one accession, i.e. $c_a > K_a$, the passing criteria was relaxed for all accessions. To do this, we allowed a disagreement between the raw base calls, $B^+{}_{Col}(p)$ and $B^-{}_{Col}(p)$ of the two strands for the Col-0 accession and between the raw base calls, $B^+{}_t(p)$ and $B^-{}_t(p)$) of the target accession. For Col-0, one of the raw base calls was allowed to differ from the reference sequence *seq(p)*, and for the target accession the raw base call of the positive strand was required to differ from *seq(p)*. Formally:

$$p_t = \Big\{ p \Big| \sum_{a=1}^{18} (\delta\{c_a(p) > K_a\}) \geq 1 \wedge \Big(B^+_{Col}(p) = seq(p) \vee B^-_{Col}(p) = seq(p) \Big) \wedge B^+_t(p) \neq seq(p) \Big\},$$

where $\delta\{.\}$ denotes the indicator function with $\delta\{true\}=1$ and $\delta\{false\}=0$. This filter further reduces the number of passing non-polymorphic sites, while retaining the majority of true SNPs (compare filter 1 to filter 2, Table S4).

Input generation, model selection, and prediction for layer 2 SVMs

For the layer 2 SVMs, we appended to the input vector $\boldsymbol{x}^{(1)}$ a binary vector b describing which of the 18 accessions passed filter 1 at the considered site p. We also included the transformed output values c from the layer 1 SVMs for all accessions (cf. Table S6):

$$\boldsymbol{x}^{(2)} = [\boldsymbol{x}^{(1)}, b, c]\cdot$$

The input vectors were again normalized on the training set. Note that both layer 1 and 2 SVMs train and predict on each accession individually. However information from multiple accessions is made available for the layer 2 SVMs. Model selection and training of the layer 2 SVMs was performed as described for layer 1 (see above). Subsequently for each position p in the 2010 dataset that passed filter 2 in accession t, exactly one prediction F^t_m was computed with the layer 2 SVM trained on accession t that had not seen the example p during training or parameter tuning. Each unlabeled position in the genome that passed filter 2 for the target accession t was predicted by all five SVMs, so that it was associated with 5 output values $F^t_1 \ldots F^t_5$. With the described cross-validation techniques we made sure that no example that had been previously used for training or model selection was used for performance evaluation. This allowed us to obtain unbiased estimates of the accuracy of our prediction methods.

Transformation of layer 2 SVM output values into confidences

As for the layer 1 SVM outputs, the corresponding outputs from layer 2 SVMs were transformed into confidence values by applying piecewise linear functions (see above). Note that the final performance is estimated on the 2010 dataset on the basis of these confidence values. However, the 2010 dataset is overrepresented for coding sequence relative to other sequence types (e.g., 2010 has 55% coding sites compared to 28% for the entire genome). Note that the sequence properties of coding sequence differ from those of other sequence types (e.g., higher GC content and lower repetitive content). Prediction algorithms are therefore likely to perform differently on the given sequence types. For this reason we determined separate transformation functions for "coding", "intergenic", and "UTR and intron" sites. Because of the comparatively small number of other site types in the 2010 dataset, we considered as "intergenic" all positions not included in a protein-coding gene model of the TAIR6 annotation (*S12*). Moreover, because of the small number of UTR sites in the 2010 dataset, we combined these with intronic sites (the ratio of UTR to intron sites is approximately the same for the 2010 dataset as for the entire genome).

The learning algorithm only classifies "no SNP" or "SNP". The final base call $B_t(p)$ for accession t at position p that corresponds to a prediction can be recovered from the intensity data, but is also subject to error (i.e., the wrong base is called at a polymorphic position). We treated these cases as false predictions. Moreover, an initial analysis of predictions revealed a high error rate at sites of exact, short, and inexact 25-mer matches (note that only dominating 25-mers were excluded by the filters). The high false call rate at these positions likely corresponds to insufficient training examples for these sites in the 2010 dataset. We therefore excluded these calls prior to the determination of piecewise linear functions and in the genome-wide predictions.

Finally, the five output values at each genomic position were transformed with the piecewise linear function corresponding to the SVM used and to the annotation of the position. We averaged over the five resulting values, thereby gaining more robust predictions.

Interpreting outputs and performance estimation

To facilitate interpretation of the predictions, we also assigned a cumulative confidence value C to each prediction (see layer 1 SVM). For instance, for all predictions having a C value greater than 0.99, a single false positive is expected for 100 predictions. The traditionally defined FDR is given by $1-C$. We have reported all predictions having a $C \geq 0.90$ (i.e., an FDR of 10%, see Section 16). We refer to this as the ML data set.

We estimated the performance of our method on the complete set of known SNPs in the 2010 dataset. As we employed cross validation and took the special composition of the labeled set into account the reported test error should generalize well to the portion of the genome that is well represented in 2010. We found that the design of the ML method leads to higher recovery for high frequency SNPs, compared to the MB method (Fig. S3).

As noted earlier, large deletions are essentially absent from 2010, and we evaluated the number of false ML calls in validated large deletions in an identical manner as for the MB predictions (see Section 5). We detected 1 false call per ~0.9 kb of deleted bases for ML predictions for $C > 0.98$. Therefore, large deleted sequences, although comparatively uncommon in the genome, are a source of additional errors that were not addressed in our analysis.

Generation of reference base calls for the ML dataset

While the ML method described above generates polymorphic base predictions, sites that are not identified as polymorphic in a given accession can be either (i) identical to the reference or (ii) polymorphic but simply not called. We used the algorithm described in Section 9 to assign base calls (either reference or "N") to positions not predicted by the ML method in a given accession but that were predicted as polymorphic with $C > 0.90$ in any other accession. For 80,087 reference base calls in this dataset represented in 2010, the rate of false assignment for reference calls was 0.049%.

7. GENERATION AND ANALYSIS OF A MERGED MB AND ML DATA SET (MBML2)

We generated a merged dataset from the MB and ML SNP predictions (MBML2) that we used for biological inferences. All MB calls were included in this dataset, and on a per-accession basis every ML call supported with an FDR of 2% was included. At positions that were included in both MB and ML calls, the rate of disagreement was 1 in 236,000. In these rare cases, an "N" was assigned as the base call.

We determined the sequence type for SNPs in MBML2 on the basis of the TAIR6 *A. thaliana* genome annotation (*S12*). "Coding", 5' and 3' untranslated regions ("UTRs"), and "intron" sequences were from the 26,541 predicted protein-coding genes. "Transposon" sequences were from gene models annotated as pseudogene and having homology to transposable elements. "Pseudogene" sequences were from gene models annotated only as pseudogene but not having strong homology to transposons. Remaining sequence was considered as "intergenic". In cases where annotations overlapped, identity was assigned with the following hierarchy: coding > UTR > intron > pseudogene > transposon > intergenic.

8. IDENTIFICATION OF HIGHLY POLYMORPHIC REGIONS

Hybridization signal on resequencing arrays is suppressed or abolished in regions of very high SNP density because successive probe sets have off-center mismatches (Fig. S1). Extended blocks of reduced hybridization signal are also expected for sequences that are deleted relative to the reference sequence. To identify such regions, we implemented a heuristic algorithm that detects extended blocks of reduced hybridization quality in a target accession relative to the Col-0 accession (i.e., background or near background hybridization). In essence, our approach identifies clusters of positions with low quality scores that are assigned with a sliding window analysis to reduce the effect of hybridization variability. Two factors confound this (or any similar) approach. First, regions harboring sequences that have poor hybridization properties have no or low hybridization to probe sets, even in the absence of polymorphic features, and can lead to false predictions. Second, cross-hybridization of repetitive sequences can mask polymorphic features. To address these issues, we excluded from the sliding-window analyses (i) positions where probe sets preformed poorly for the Col-0 reference, and (ii) positions with exact, short, or inexact 25-mer matches elsewhere in the genome.

Assigning scores to informative positions

Two scores, $\bar{s}_{QR}$ and $\bar{s}_{MM}$, were used as indicators for highly polymorphic sequence tracts. Initially, we calculated a value $s_{QR}(p)$ for each non-repetitive position p as follows:

$$s_{QR}(p)=\begin{cases}\dfrac{n}{Q_t^+(p)+Q_t^-(p)} & if \quad n>6\\ 0 & else\end{cases}$$

$$\text{with } n=Q_{Col}^+(p)+Q_{Col}^-(p),$$

where $Q_{Col}^+(p)$ and $Q_{Col}^-(p)$ are the quality scores at position p of the Col-0 accession for the forward and reverse strand respectively, and similarly $Q_t^+(p)$ and $Q_t^-(p)$ for the target accession t. A high value of s_{QR}, indicating a high probability for being polymorphic, results at positions p where the target accession has low quality scores relative to the reference. At positions where the sum of both quality scores for Col-0 was ≤ 6 (i.e., low/unreliable hybridization), s_{QR} was set to 0.

Subsequently, values of s_{QR} were used in a sliding window analysis to assign to each position p with $s(p)\neq 0$ the quality ratio score $\bar{s}_{QR}$. This ratio score is defined as:

$$\bar{s}_{QR}(p)=quart\{s_{QR}(p')\mid p'\in w\}.$$

Here w is a window centered on p for which contiguous positions are included on either side of p following the removal of all repetitive positions (positions with exact, short, or inexact 25-mer matches) and positions for which $s_{QR}=0$. By visual inspection the 1st quartile (*quart*) was found to preserve sharp transitions (e.g., at deletion breakpoints).

The second score, $\bar{s}_{MM}(p)$, is defined as the difference between the number of mismatch calls $[B_t^{str}(p) \neq seq(p)]$ on both strands, $str \in \{+,-\}$, for the target accession t and the Col-0 accession $[B_{Col}^{str}(p) \neq seq(p)]$ within the window w, normalized by the length of the window:

$$\bar{s}_{MM}(p) = \frac{1}{|w|}\left(\sum_{str=\{+,-\}} \sum_{p' \in w} M_t^{str}(p') - \sum_{str=\{+,-\}} \sum_{p' \in w} M_{Col}^{str}(p') \right)$$

where $M_t^+(p) = 1$ if $B_t^+(p) \neq seq(p)$ and else $M_t^+(p) = 0$ and similar for $M_t^-(p)$, $M_{Col}^+(p)$ and $M_{Col}^-(p)$.

The extent to which these scores discriminate between deleted and present sequences is shown for one accession, Br-0, for w = 101 (Fig. S9). Positions covered by sequence data from the 2010 fragments were partitioned according to their score and the abundance of SNPs, conserved regions, and deletions. The overlapping distributions indicate the limits of sensitivity and specificity. Longer deletions can be detected more easily than shorter ones.

Generating Polymorphic Region Predictions (PRPs)

In a first step, we identified positions for inclusion in PRPs where both (i) $\bar{s}_{QR}(p)$ was above threshold t_{QR} and (ii) $\bar{s}_{MM}(p)$ was above threshold t_{MM}. Secondly, we clustered positive sites by determining regions of ≥ 50 positive sites for which gaps of ≤ 10 negative sites were tolerated. Clusters of positive sites meeting this requirement were designated as PRP cores, and corresponded to a set of conservative initial predictions. However, larger polymorphic or deleted regions may contain several such initial predictions. Thus, adjacent cores were merged if the region in between was also likely to be deleted or highly polymorphic. As merging criteria $s_{merge,sc}$ we defined the following for the two scores $sc \in \{QR, MM\}$:

$$s_{merge,sc} = \frac{|C_1|\, t_{sc} + |C_2|\, t_{sc}}{|G|\, t_{sc} - \sum_{p \in G} \min(t_{sc}, \bar{s}_{sc}(p))}$$

Here $|C_1|$ is the length of the first core C_1 (similarly for C_2) and $|G|$ is the length of the gap between the two cores. Both values, $s_{merge,QR}$ and $s_{merge,MM}$ had to be ≥2 for core merging. Fig. S10A shows a representation of this formula; with green areas corresponding to the numerator and the red area to the denominator.

Given a core prediction we then estimated the closest positions upstream and downstream for which hybridization resembled the reference. In case of a deletion polymorphism, this amounts to predicting intervals (i.e., boundaries) in which the breakpoints reside. The sites closest to the core at which both scores fell below a second pair of thresholds (u_{QR} and u_{MM}) were taken as initial end points. The initial boundary estimation was then refined with an iterative procedure (see Fig. S10B,C). To delineate the boundary regions more precisely, in each step the window size was reduced by 20% and in the boundary region deletion scores were re-computed. If – by intersection with the score thresholds – a new boundary interval was completely contained in the original boundary, the boundary was shortened, thereby extending

the core. This step was repeated as long as determining a new boundary interval was possible and the window size was at least 5 bp. Determining a new boundary interval was considered impossible and boundary refinement was terminated when there were two or more possible new boundary intervals which did not overlap. (In case of overlapping intervals the smallest one, which is contained in all larger intervals, was chosen as new boundary and boundary refinement was continued.)

In a final step of boundary refinement, we checked whether boundaries contained contiguous stretches where hybridization of reference probes produced higher intensities than non-reference probes. We call these contiguous stretches *conserved words*. We expected the length of conserved words to be smaller in highly polymorphic regions compared to conserved regions and therefore truncated boundaries if they contained long conserved words close to their end points. We proceeded as follows. First, the core was extended into the boundaries until a conserved word of length ≥ 6 or nearby conserved words of length $n \in \{3,4,5\}$ at a distance of $\leq n^2$ to each other were encountered. Second, the boundaries were truncated at the outer end such that conserved words of length $n \geq 5$ within a distance to the previous endpoint of $\leq n^3$ were excluded. Third, if a boundary had not been truncated in the second step, it was extended until either a conserved word of length ≥ 10 was encountered or nearby conserved words of length $n \geq 5$ within a distance of $\leq n^2$ were encountered.

Finally, in a few cases PRPs overlapped (PRPs were generated independently). Where cores overlapped, PRPs were always merged; if only the boundaries overlapped, we used the same formula as for core merging, but this time the two ratios $s_{merge,sc}$ had to be ≥5. If predictions could not be merged by these criteria, they were discarded.

Choice of thresholding parameters for genome-wide predictions

For the recognition of sites in deletions (for deletions ≥ 25 bp in the 2010 dataset), we determined the dependency of sensitivity and specificity on the threshold values t_{QR} and t_{MM}. Fig. S9 shows this dependency for accession Br-0. Across all accessions, the mismatch score was observed to be more robust than the quality ratio score. On the basis of data presented in Fig. S9, we chose a threshold value of 0.72 for the mismatch score and a value of 3.8 for the quality ratio score (*w* was set to 101 throughout). The lower thresholds u_{QR} and u_{MM} were adjusted by visual inspection of the surrounding regions of several long (>25) deletions in non-repetitive regions of the 2010 dataset. A threshold value of 2.5 was chosen for u_{QR} and 0.32 for u_{MM}. The number of PRPs generated per accession with these parameters is given in Table S7.

PRP-based analyses

While the PRPs consist of "core" and "boundary" regions, unless otherwise noted, all analyses are based on the core portion of PRPs generated with the most stringent criteria. We used boundary information to facilitate experimental validation of PRPs (see Section 11), and we release the boundary information to facilitate experimental studies by the scientific community (see Section 16).

9. PREDICTION OF NONPOLYMORPHIC BASES

We implemented a thresholding algorithm to assign reference base calls to nonpolymorphic positions interrogated with the arrays. The approach is motivated by the observation that while SNP and deletion features cause extended regions of low quality scores, positions with low quality scores (e.g., at positions with poorly performing probe sets) embedded in regions with high quality scores and for which maximal intensities match the reference sequence are unlikely to be polymorphic. The base calling algorithm assigns a call *C(p)* to each non-repetitive position *p* in the genome, which is either the reference base call $seq(p)$ or an ambiguous call "N". It checks the following conditions until *C(p)* is assigned. By $s = \arg\max_{r \in \{+,-\}} Q^r(p)$ we denote the strand with the higher quality score:

CONDITION 1:
If each position in window *w* centered on *p* is non-repetitive,
    check condition 2.
Else:
    if $\left(B^+(p) = B^-(p) = seq(p)\right) \wedge \left(Q^s(p) \geq t_1\right)$,
        set $C(p) = seq(p)$, done.
    else set $C(p) = \mathbf{N}$, done.

CONDITION 2:
If $\left(B^s(p) = seq(p)\right) \wedge \left(Q^s(p) \geq t_2\right)$,
    check condition 3.
Else set $C(p) = \mathbf{N}$, done.

CONDITION 3:
Determine a set of positions *P(p)* in the window *w*:
$P(p) = \left\{P : \left(B^s(P) = seq(P)\right) \wedge \left(Q^s(P) \geq t_2\right)\right\}$.
If $|P(p)| \geq t_3$,
    check condition 4.
Else set $C(p) = \mathbf{N}$, done.

CONDITION 4:
If $\underset{p' \in P}{\text{mean}}\left(Q^s(p')\right) \geq t_4$,
    set $C(p) = seq(p)$, done.
Else set $C(p) = \mathbf{N}$, done.

The parameters *w*, t_1, t_2, t_3, and t_4 can be adjusted to control precision and recall. On the basis of inspection of quality score information for the first 3,000 positions of chromosome 1 from the Col-0 reference accession, we set these parameters to 7, 20, 7, 6, and 10 for calling reference bases for all accessions (including the Col-0 reference itself). The number of bases predicted as reference per accession with these parameters is given in Table S8.

Performance and evaluation
Performance was evaluated against the 2010 dataset by summing over accessions. For positions with a substitution in another accession, 66% of known reference bases were assigned as reference by the base calling algorithm (on the basis of 170,386 examples). In contrast, the corresponding rate of false reference base assignment was 0.46% (on the basis of 48,692 examples). In addition, we determined the number of reference bases predicted in known deletions (see Sections 5 and 11), and observed 1 false reference prediction per 71 known deleted positions visible to the base calling algorithm. In addition to experimental variability, several factors likely account for the false reference base calls. First, in making reference base calls, we only filtered positions for exact, short, and inexact 25-mer matches. Nevertheless, probes with multiple mismatches, or small indels, may still cross-hybridize and lead to false predictions. Second, our correction for repetitive probe sets was necessarily based on the reference sequence from Col-0, and does not correct for unidentified repetitive sequences that may be present in a given target accession.

Construction of pseudochromosome sequences
To facilitate use of our dataset by the scientific community, we generated pseudochromosome sequences for each of the 20 accessions (see Section 16). To construct the pseudochromosome sequences, reference base calls were from the above described algorithm, while SNPs were from MBML2. In the pseudochromosome sequences, ambiguous positions are denoted by an "N", while repetitive positions that were masked are denoted as "R".

10. EFFECTS OF SNPs ON GENE MODELS

We assessed the effects of SNPs in the MBML2 dataset on the 26,541 nuclear protein-coding gene models for the TAIR6 release of the *A. thaliana* Col-0 genome annotation (*S12*). Effects were assessed on a per accession basis, and the reference sequence was used for base assignment at positions not predicted to be polymorphic in MBML2. Where more than one isoform for a gene was annotated, effects were determined on an isoform basis. Absolute numbers for "large-effect SNPs" are given in Table S9. We defined a large-effect SNPs as (i) introducing a premature stop codon, (ii) changing a stop codon in the reference to coding potential, (iii) generating a nonfunctional splice donor site, (iv) generating a nonfunctional splice acceptor site, or (v) disrupting an initiation methionine codon. Although not considered as large-effect SNPs, substitutions converting consensus splice donor sites to nonconsensus sites (GT to GC) or vice versa were also assessed (Table S9). The effect of these changes on splicing is expected to vary depending on sequence context (*S13*).

11. VALIDATION OF LARGE-EFFECT SNPs AND PRPs

Verification of large-effect SNPs
A subset of large-effect SNPs supported by the MB method was characterized by PCR and dideoxy sequencing with flanking primers. For validation, SNPs were selected randomly with respect to predicted biological effect and gene category, and validated from a single accession. To match accessions to predictions for validation, an accession harboring a given prediction was

chosen at random from all accessions predicted to share the same substitution. From this list of accessions and predictions, we attempted to validate all predictions from accessions Bay-0, Bor-4, Br-0, and Bur-0. In addition, we attempted to validate a minimum of 44 predictions, ordered by chromosome 1-5 and position, from each of the remaining accessions.

Primer pairs used for prediction verification were synthesized on a Genemachines Polyplex oligosynthesizer, and were designed with the program Primer3 (*S14, S15*) to be a minimum of 150 bp from the predicted SNP, to amplify an ~500 bp product, and to have a T_m of ~58˚C and GC content between 40 to 70%. PCR was performed in 15 μl reactions with 10 ng of genomic DNA, 1.25 U Taq polymerase, and final concentrations of 50 mM KCl, 10 mM Tris-HCl pH 8.3, 1.5 mM $MgCl_2$, 0.2 mM dNTPs, and 0.2 μM each primer. For amplification, reactions were heated to 94˚C for 2 minutes, followed by 30 cycles of 94˚C for 0.5 minutes, 55˚C for 0.5 minutes, 68˚C for 1 minute, with a final 5 minutes at 68˚C.

Where PCR product was detected by gel electrophoresis, dideoxy sequencing was performed with either the forward or reverse primer used for amplification. For sequencing, 13 μl of each reaction was added to 0.04 μl Exonuclease I (Fermentas, 20U/ul), 0.8 μl shrimp alkaline phosphatase at 1 U/μl (New England Biolabs, Ipswich, MA), and 3.16 μl sterile water. The resulting mixture was incubated at 37˚C for 45 min to degrade excess primers and nucleotides from the amplification step, followed by 80˚C for 10 min to inactivate enzymes. Following the addition of 20 μl of water to each sample, 2 μl was used in a sequencing reaction containing 2 μl 5X sequencing buffer (Amersham, Piscataway, NJ), 0.5 μl primer (20 μM stock), 2 μl sterile water, and 1 μl Amersham ET Terminator mix. Cycle sequencing was performed with 25 repetitions of 95˚C for 0.2 min and 60˚C for 1 min. Sequencing reactions were sodium acetate/ethanol precipitated, resuspended in 10 μl water, and analyzed on a ABI 3700 sequencing machine (Applied Biosystems, Foster City, CA).

A given sequence read was aligned against the corresponding sequence from the requisite accession (the accession-specific SNP predictions and up to 500 bp of flanking sequence from the Col-0 reference) with BLASTN 2.2.2 (*S16, S17*). From the resulting alignments, we identified the base in the dideoxy sequence read corresponding to the position for which the large-effect SNP was predicted. We also determined the *Phred* (*S5, S18*) quality score that corresponded to the position. We employed a Perl script to perform these tasks. Where verification attempts failed at the PCR or sequencing steps, or where the *Phred* quality score at the base targeted for validation was < 20, attempts were considered as unsuccessful (Table S9). Successful validation attempts are reported in Table S10.

In addition, for predictions affecting coding sequences (i.e., premature stop codons), we inspected the nearest 2 bp that flanked the predicted large-effect SNP. For 3 substitutions predicted to introduce premature stop codons, a flanking nucleotide substitution was detected by dideoxy sequencing that was not predicted from the array data, and that together with the predicted SNP generated a missense alteration as opposed to a premature stop change (2 substitutions in the same codon). These instances are excluded from Table S10.

Characterization of PRPs corresponding to deleted sequences

We analyzed a subset of PRPs with PCR and sequencing strategies similar to that employed for large-effect SNP validation. We chose PRPs where the length of the core prediction was ≥ 300 bp, the flanking boundary predictions were ≤ 100 bp, and the core overlapped the coding sequence of one or more gene models. Where multiple PRPs overlapped the coding sequence for a single gene, either within the same accession or among accessions, a single PRP was chosen at

random. Subject to these criteria, PRPs were selected randomly with respect to genomic location.

Primers used for amplification were chosen ~250 bp from PRP boundaries such that the expected size of amplicons would be ~500 bp under the assumption that an entire PRP (core plus boundary regions) corresponds to a deletion relative to the reference genome sequence. A caveat of this approach is that non-deletion PRPs will fail to amplify if longer than one or two kb.

Where products could be amplified, sequencing was attempted with both the forward and reverse amplification primers. For deletion/polymorphism detection, sequence reads were trimmed with the *Pregap4* program in the Staden package (*S19, S20*) with window length set to 50 bp and mean *Phred* score of ≥ 20. We next determined the best match for both the forward and reverse strand reads against the entire reference genome sequence with BLASTN to detect spurious/nonspecific amplification (e.g., amplification of repetitive sequences). If the highest matching genomic hit was not coincident with the target PRP coordinates or many hits were observed, the verification attempt was considered as negative. Forward and reverse strand contigs were next assembled with *Gap4* from the Staden package for instances where reads overlapped. The consensus, forward, and reverse reads for a given prediction were then aligned to the target Col-0 reference sequence (the entire sequence between primer pairs used for amplification) with the program MUSCLE (*S21, S22*), and alignments were subsequently manually curated. In some cases, sequence was available from only the forward or the reverse reads, or the forward or reverse reads did not overlap. Where deletions or stretches of polymorphisms were detected for these partial alignments, a given PRP was considered as verified. Otherwise, the attempt was considered to have failed and the given sequence alignment to be incomplete.

For instances where deletions of ≥ 50 bp were validated, the relationship to gene models was assessed (Table S11). In other cases, PRPs corresponded to clusters of SNPs and small indels, and drastic effects on gene models could be inferred in some cases (e.g., 1 bp indels introducing frameshift mutations). However, many PRPs were extremely polymorphic and could not be unambiguously aligned, or were supported by single strand reads (i.e., only the forward or reserve read; see also Table S11). In these cases, additional sequencing is required to fully characterize effects on gene models.

12. ANALYSIS OF POLYMORPHISMS BY GENE CATEGORIES

We assessed the distribution of "major-effect changes" by gene category. Major-effect changes were defined to include large-effect SNPs and PRP overlaps to coding sequences.

Gene categories were constructed as follows. Annotation status: Expression support was given to the 26,541 annotated *A. thaliana* coding genes on the basis of full-length cDNAs, ESTs, MPSS, SAGE, and genome-wide tiling array transcriptome evidence (*S23-S29*). Genes without evidence of expression were assigned as "not expressed". Otherwise, genes that were expressed by our criteria that had been annotated as, for example, "expressed" or "hypothetical" in the TAIR6 annotation, were denoted "expressed unknown". All other genes were assigned as "expressed known". We note, however, that some assignments between "expressed unknown" and "expressed known" are potentially incorrect or are ambiguous, in part as a result of inconsistencies in the existing annotation. Duplication status: Assignment as segmental or tandem duplicates as were Haas *et al.* (*S30*) (genes annotated as both segmental and tandem duplicates were excluded from analysis). Gene family status: Gene family or superfamily lists

were from TAIR (*S12*), Shiu and Bleecker (*S31*) (receptor-like kinase genes), Meyers *et al.* (*S32*) (NB-LRR genes), or as provided by R. Vierstra, D. Gingerich, and J. Gagne (Univ. of Wisconsin; F-box genes). For gene family analysis with NB-LRR genes, we included members with complete TIR-NBS-LRR or CC-NBS-LRR domain structure and open reading frames as annotated for the reference sequence. Homology to poplar: A list of *A. thaliana* genes with no or low homology to genes in poplar was provided by L. Sterck and Y. van de Peer [see also (*S33*)]. Gene numbers reported for the various categories differ from that reported in source gene lists for several reasons. First, outdated gene models (no longer present in the TAIR6 annotation) were dropped. Second, for our analyses of major-effect changes, we excluded genes that were entirely repetitive (i.e., every position corresponded to exact or short 25-mer matches), and for which no SNP predictions could be generated by our algorithms.

In addition to counting genes per category with major-effect changes (Fig. 3), we normalized large-effect SNPs by the number of non-repetitive sites for all genes in a given category (Fig. S11A). A related normalization was performed for PRPs by gene category (Fig. S11B).

13. ALLELE FREQUENCY ANALYSIS FOR SNPs IN CODING SEQUENCES

For allele frequency analyses, we excluded SNP positions where in any target accession polymorphisms were present within 2 bp. This allowed unambiguous assignment of synonymous and nonsynonymous sites, and also removed nearly adjacent SNPs for which the rate of false prediction is expected to be highest (e.g., see Fig. 1E). We also limited our analysis to diallelic SNPs. For consistency, large-effect SNPs were selected for inclusion in allele frequency analyses with the same criteria. The occurrence of the minor allele was determined by subsampling at positions for which at least 16 calls were generated. For such positions, 16 calls were selected at random to determining the occurrence of the minor allele. Supplemental analysis for allele frequency by gene family is given in Fig. S12.

14. GENOME-WIDE PATTERNS OF POLYMORPHISM

Nucleotide diversity for the set of 19 accessions (excluding Van-0) was estimated from the pseudochromosome sequence (see Section 9) by averaging the differences per base (total mismatches divided by total base comparisons) for a particular class of sites across all pairs of accessions (Figs. 4 and S13-S14). At each site, comparisons were between pairs of accessions that were not called "N" or "R". To estimate nucleotide diversity for different classes of sites (e.g. intergenic or four-fold degenerate protein coding), only those sites were used in comparisons, though window sizes were defined according to absolute distance along the reference sequence. In each window, a minimum number of base comparisons were required between each pair of accessions for that pair to contribute to the average pairwise diversity. For diversity at four-fold degenerate sites, at least 100 base comparisons per 50 kb window were required between each pair of accessions, at least 250 base comparisons per 250 kb window, and at least 500 base comparisons per 500 kb window were required. For diversity at intergenic sites, at least 2,000 base comparisons per 50 kb window were required between each pair of accessions, and at least 5,000 base comparisons per 250 kb window.

For the 2010 dataset, nucleotide diversity was estimated for 95 accessions (Van-0 excluded) from four-fold degenerate coding sites from 1,051 public sequence fragments with at least one four-fold degenerate coding site. The majority of these fragments were described in Nordborg *et al.* (*S7*), and the others are available for download (see Section 16). These fragments are nearly identical to the 2010 fragments described in Section 5. All heterozygous sites and deletion sites were treated as missing data to make the estimates more comparable to that for the array data (which uses only SNP calls). Otherwise, nucleotide diversity was estimated as in the preceding paragraph. Nucleotide diversity was estimated in windows of 500 kb, and only a single base comparison was required between each pair of accessions for that pair to contribute to average pairwise diversity.

Correlations of nucleotide diversity estimates with several genomic factors were explored in windows of 50 kb (Table S12). These included: number of NB-LRR genes, number of all genes (excluding pseudogenes), number of repetitive probes, distance to the centromere, GC content from the Col-0 pseudochromosome, and amount of missing data (both repetitive probes and sites where a confident call could not be made). The number of NB-LRR genes was obtained by counting the number of these genes that overlap with each 50kb window [NB-LRR genes included in this analysis are or have homology to TIR-NBS-LRR or CC-NBS-LRR genes, and were collected from Meyers *et al.* (*S32*) and the TAIR6 gene annotation (*S12*)]. Number of total genes was counted from the number of gene "midpoints" (average of gene model start and end, TAIR6 gene annotation in each window). Number of repetitive probes was the count of the positions masked as "R" in the pseudochromosome sequence (see Section 9) for either intergenic sites, four-fold degenerate coding sites, or all types of sites. To estimate distance from centromeres, centromeres were heuristically defined as the span of 50 kb windows such that outside of centromeres no runs of 5 consecutive windows exist where the proportion of repetitive probes is >40% in each of the 5 windows. This produced centromeres between 13.7–15.9 Mb for chromosome 1, 2.45–5.5 Mb for chromosome 2, 11.3–14.3Mb for chromosome 3, 1.8–5.15Mb for chromosome 4 (a span that includes the knob and inversion on the top of this chromosome), and 11–13.35Mb for chromosome 5. Distance from centromeres was 0 for windows within centromeres thus defined but was otherwise the shortest distance from the edge of the centromere and the edge of the window. GC content was measured as the number of sites called "G" or "C" divided by the number of sites called "G", "C", "A", or "T" in the Col-0 pseudochromosome for either intergenic sites, four-fold degenerate coding sites, all coding sites, or all types of sites. The amount of missing data was measured as the proportion of all sites called either "N" or "R" averaged over the pseudochromosomes from 19 accessions (Van-0 excluded) at either intergenic sites, four-fold degenerate sites, or all types of sites. Multiple regression analyses with stepwise model selection were performed with the statistical package R (*S34*). The relationship between intergenic nucleotide diversity and the following predictor variables was investigated: number of NB-LRR genes, number of all genes, distance to the centromere, and the following three factors measured at both intergenic and all sites - number of repetitive probes, GC content, and missing data. A similar analysis was performed for four-fold degenerate nucleotide diversity, with four-fold degenerate sites instead of intergenic sites and the addition of GC content at all coding sites added as another predictor variable.

15. SCANNING FOR RECENT SELECTIVE SWEEPS

We examined the extent of haplotype sharing among accessions to identify candidate regions for selective sweeps. To do this, we split the genome into non-overlapping 10 kb windows and calculated the proportion of differences between all pairs of accessions in each window. For a site to factor into this calculation, neither member in a pair of accessions being compared could have missing data. Then all runs of five or more consecutive 10 kb windows that each had fewer than 1 difference per 1,000 comparisons were identified. When a 10 kb window had more than 90% missing data for a pair, this window was not counted towards the minimum five windows required; it was, however, allowed to extend a run. The resulting runs are shown in Fig. 5 for chromosome 1, and in Figs. S17 and S19 for chromosomes 1-5.

To identify the best candidates for recent partial or complete sweeps, we determined, for each 10 kb window, the total length of runs that include this window across all accession pairs. The highest total run lengths can represent regions where almost all accession pairs are highly similar, but may also include regions where fewer pairs are similar but over much longer runs. An alternative method to identify candidates for sweeps is simply to count for each window the number of pairs of accessions with a run overlapping this window out of a maximum possible of 171 (from 19 accessions, Van-0 excluded). This approach can identify short complete sweeps or short deep partial sweeps missed by the previous approach, but generally does not distinguish between similarity across the minimum 50 kb distance versus much more extensive similarity. The results of both approaches are shown in Fig. S18.

16. DATA RELEASE

We have deposited processed resequencing data in the NCBI Trace Archive (*S35*). Each trace file represents data for one contiguous fragment of tiled sequence in one orientation. The trace amplitude data consists of mean fluorescence intensity measurements for each feature on the array. The called sequence consists of the brightest of the four nucleotide probes for each position in the reference sequence. Data for the reverse tiling is reverse complemented before the trace files are generated, so that the forward (A) and reverse (Z) reads are both reported for the "+" strand of the reference sequence. In addition to the basic experimental data, called sequence, and quality scores, each trace also carries descriptive information, the structure of which is specified by the NCBI Trace Archive. Table S15 explains how to interpret some of these fields for Perlegen resequencing traces, and supplements the Trace Archive documentation.

Additional data is hosted at TAIR (*S12*). Included are comma or tab delimited files specifying all SNPs and PRPs, effects of SNPs on coding gene models (both nonsynonymous and synonymous SNPs are annotated), an annotation of core PRP overlaps to coding genes, and pseudochromosome sequences for each accession. For the SNP annotation, inclusion in MBML2 is indicated, and probability values for the ML method are given. SNPs determined to be incorrect by dideoxy sequencing (Table S10) have been removed from the release. This results in small differences in SNP numbers relative to that reported for predicted SNPs elsewhere in the manuscript (e.g., Table S3). A list of the 26,541 coding genes annotated by the categories used for constructing Fig. 3 has also been provided, and the occurrence of all major-effect changes by gene is summarized in the same file with information about verification where available (see Tables S10 and S11). The data in Tables S10 and S11 are also provided as text files at TAIR. A

list of dideoxy validated deletions (and other polymorphism types, such as insertions) discovered during PRP validation attempts is also available. The coordinates for all polymorphisms are given by chromosome and position [on the basis of (*S1*)]. Sequence data for large-effect SNP and PRP validations (Section 11) have been deposited in GenBank (EI100660- EI102044). In addition, we have provided as alignments all sequence data for the 2010 dataset that was used in this study.

SUPPORTING REFERENCES AND NOTES

S1. TIGR genome assembly version 5.0 NCBI *Arabidopsis thaliana* repository.
S2. M. Chee *et al.*, *Science* **274**, 610 (1996).
S3. D. A. Hinds *et al.*, *Science* **307**, 1072 (2005).
S4. N. Patil *et al.*, *Science* **294**, 1719 (2001).
S5. B. Ewing, P. Green, *Genome Res.* **8**, 186 (1998).
S6. I. Lee, A. A. Dombkowski, B. D. Athey, *Nucleic Acids Res.* **32**, 681 (2004).
S7. M. Nordborg *et al.*, *PLoS Biol.* **3**, e196 (2005).
S8. B. Schölkopf, A. Smola, *Learning with Kernels* (MIT-Press Cambridge, 2002).
S9. V. N. Vapnik, *The Nature of Statistical Learning Theory* (Springer, 1995).
S10. S. Sonnenburg, G. Rätsch, C. Schaefer, B. Schölkopf, *Journal of Machine Learning Research*, 1531 (2006).
S11. R. O. Duda, P. E. Hart, D. G. Stork, *Pattern Classification, 2nd ed.* (Wiley & Sons, Inc., New York, 2000).
S12. The Arabidopsis Information Resource (http://www.arabidopsis.org/).
S13. G. Rätsch, S. Sonnenburg, *Accurate Splice Site Prediction for C. elegans*. B. Schölkopf, K. Tsuda, J. P. Vert, Eds., Kernel Methods in Computation Biology (MIT Press, Cambridge, MA, 2004).
S14. http://fokker.wi.mit.edu/primer3/.
S15. S. Rozen, H. Skaletsky, *Methods Mol. Biol.* **132**, 365 (2000).
S16. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J. Mol. Biol.* **215**, 403 (1990).
S17. http://www.ncbi.nlm.nih.gov.
S18. http://www.phrap.org/.
S19. http://www.sanger.ac.uk/Software/production/staden/.
S20. R. Staden, *Mol. Biotechnol.* **5**, 233 (1996).
S21. R. C. Edgar, *BMC Bioinformatics* **5**, 113 (2004).
S22. R. C. Edgar, *Nucleic Acids Res.* **32**, 1792 (2004).
S23. M. S. Boguski, T. M. Lowe, C. M. Tolstoshev, *Nat. Genet.* **4**, 332 (1993).
S24. Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/projects/geo/).
S25. C. Lu *et al.*, *Science* **309**, 1567 (2005).
S26. M. Seki *et al.*, *Science* **296**, 141 (2002).
S27. K. Yamada *et al.*, *Science* **302**, 842 (2003).
S28. Arabidopsis Transcriptome Express Tool (http://signal.salk.edu/cgi-bin/atta).
S29. Arabidopsis Unannotated Secreted Peptide Database (http://peptidome.missouri.edu/).
S30. B. J. Haas *et al.*, *BMC Biol.* **3**, 7 (2005).
S31. S. H. Shiu, A. B. Bleecker, *Plant Physiol.* **132**, 530 (2003).
S32. B. C. Meyers, A. Kozik, A. Griego, H. Kuang, R. W. Michelmore, *Plant Cell* **15**, 809 (2003).
S33. G. A. Tuskan *et al.*, *Science* **313**, 1596 (2006).
S34. R. Ihaka, R. Gentleman, *J. Comput. Graph. Stat.* **5**, 299 (1996).
S35. The NCBI Trace Archive (http://www.ncbi.nlm.nih.gov/Traces/trace.cgi).
S36. C. Toomajian *et al.*, *PLoS Biol.* **4**, e137 (2006).

SUPPORTING FIGURES

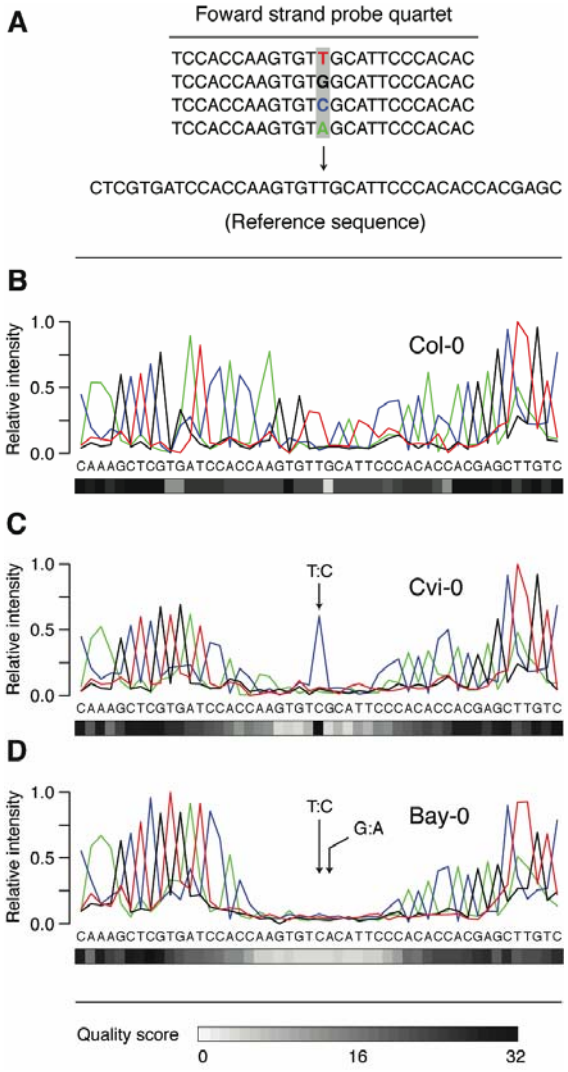

Figure S1. Experimental design and polymorphic signatures. (A) Each forward and reverse base was queried with a probe quartet. (B-D) Pseudotrace representations for Col-0 (the reference sequence), Cvi-0, and Bay-0 for a region on chromosome 1. Peaks correspond to normalized intensities for forward strand probe quartets. Known sequence and quality scores are shown beneath each trace. Closely linked SNPs (D) suppress SNP signatures, because none of the alternative probes is without mismatch.

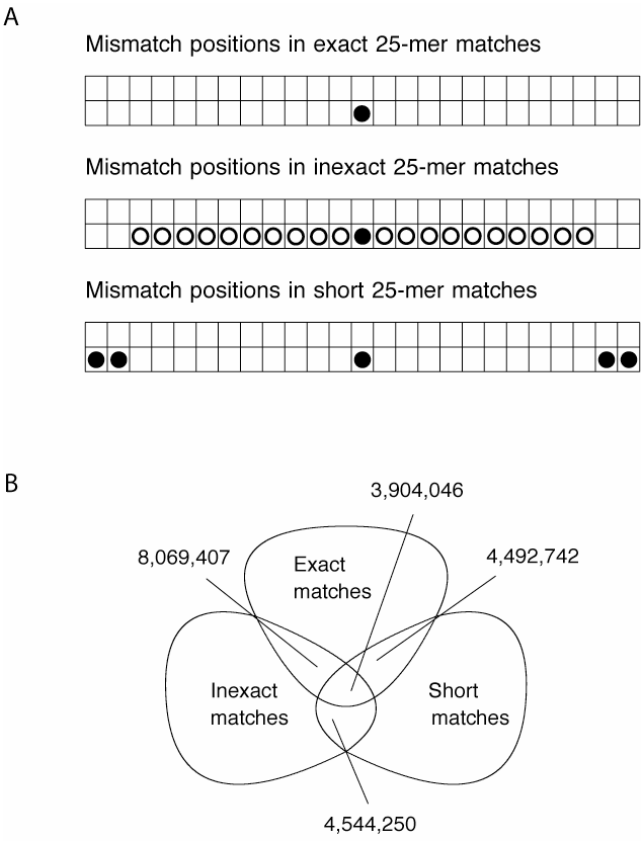

Figure S2. Match type definition for 25-mers and nonredundant overlap of match types. (A) Positions at which mismatches are tolerated in the three 25-mer match types. Squares denote positions in probes from 1 to 25, and filled circles indicate positions for which mismatches are tolerated. For inexact matches, a single mismatch at one of the positions indicated by open circles is tolerated. (B) Intersection between non-redundant positions with k-mer matches. For example, of 8,069,407 positions where there is an exact and inexact 25-mer match, 3,904,046 also have a short 25-mer match. Absolute numbers for match types are given in Table S2.

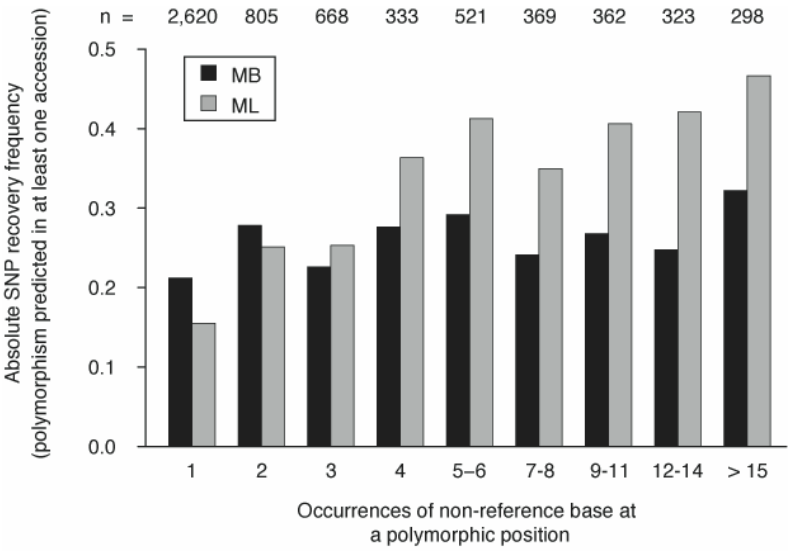

Figure S3. Recall by position as a function of occurrence of the non-reference base (assessed against the 2010 dataset when complete data was available). Use of information across accessions by the ML method leads to enhanced recall for substitutions that are present at moderate to high allele frequencies relative to the reference base at a position. Recovery by the MB method as a function of allele frequency was determined to be similar.

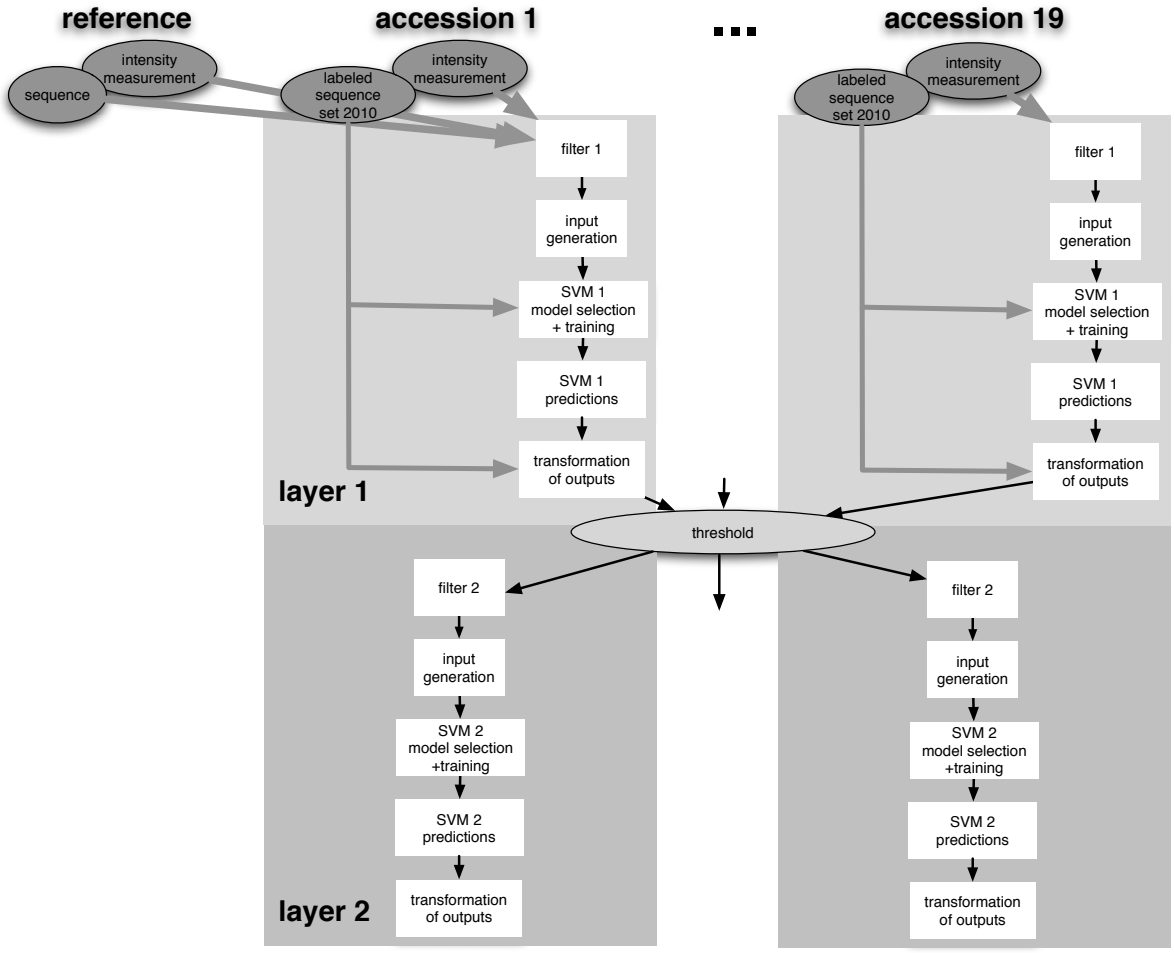

Figure S4. Flow chart describing the two-layered machine learning approach to SNP calling. In layer 1, only data from the target and reference accessions were used. Information across all accessions was exploited in a second step, in layer 2.

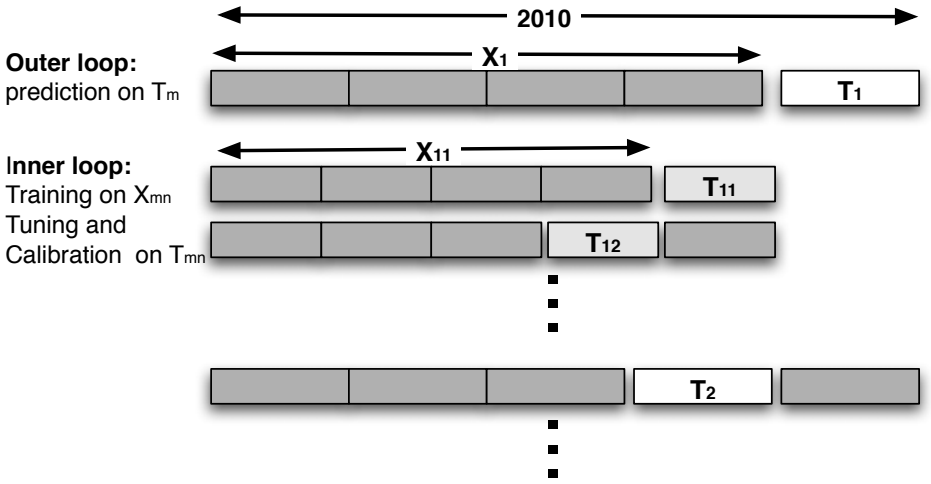

Figure S5. Methods of cross-validation scheme for SVM training and evaluation. We performed 5-fold cross validation to predict each position of the labeled set with an SVM that had not seen the example during training or parameter tuning. During model selection, k different models were trained on each subset X_{mn}. Parameter settings that performed best on the set T_{mn} were selected. The performance of each of the five SVMs was tested on the corresponding subset T_m. The subset T_m was also used to estimate the transformation of SVM output values to confidence values.

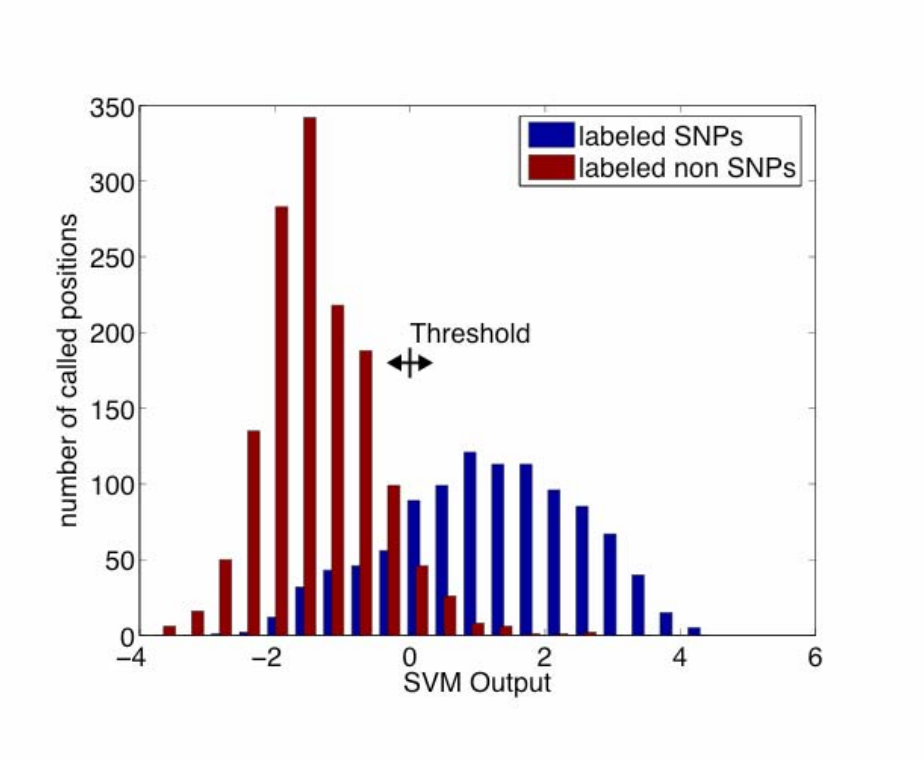

Figure S6. Histogramm of outputs from ML algorithm for SNP and non-SNP positions. By shifting a threshold on the output values, the number of called sites can be adjusted with respect to false positive SNPs. Each threshold therefore corresponds to a number of true positives (TP) and false positives (FP).

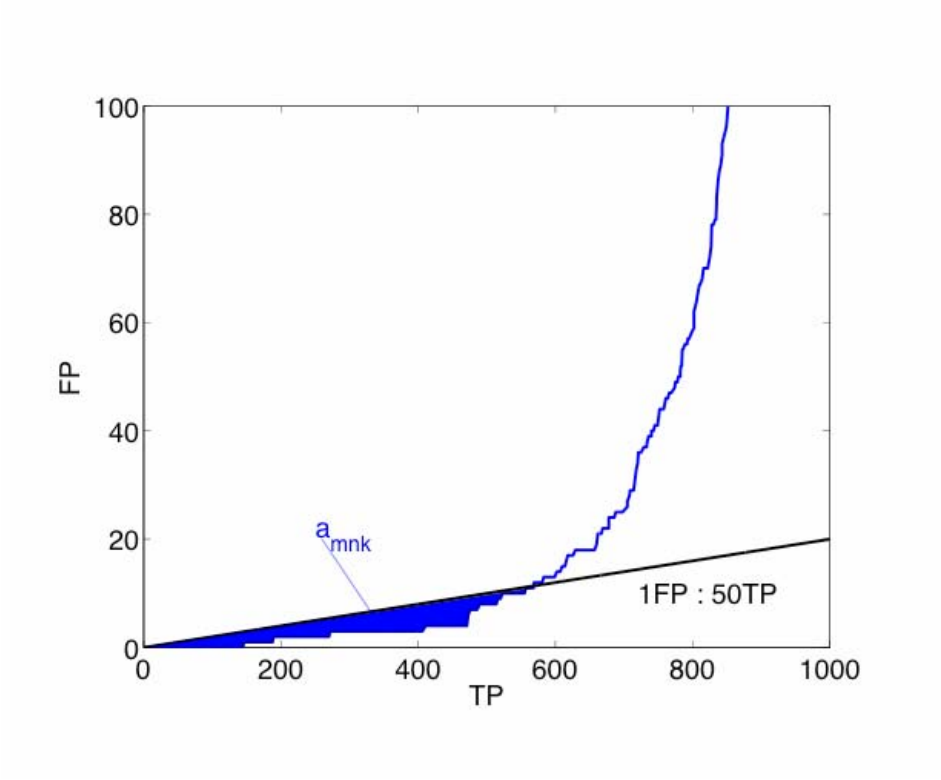

Figure S7. The performance of the SNP calling approach was optimized not only on a single point on a receiver operating characteristic (ROC) curve, but over whole range of low false discovery rates. We chose the model k which maximized the area a_{mnk} between the computed curve $FP=FP(TP)$ and a line representing 1 FP at 50 TP. This measure also proved to be more stable than a single point.

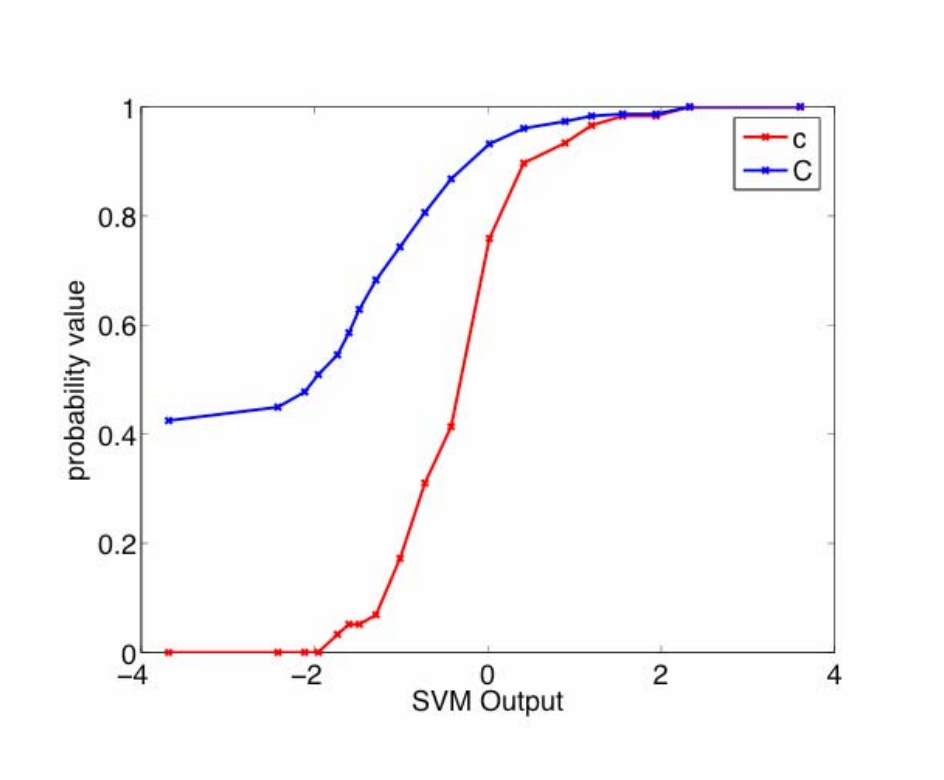

Figure S8. For each trained SVM we determined piecewise linear functions on the subsets T_m. SVM output values are thereby mapped to probability values c, reflecting the likelihood of a true positive for any specific prediction. We additionally defined a cumulative probability C, which describes the likelihood for any prediction with $c \geq C$ to be a true positive.

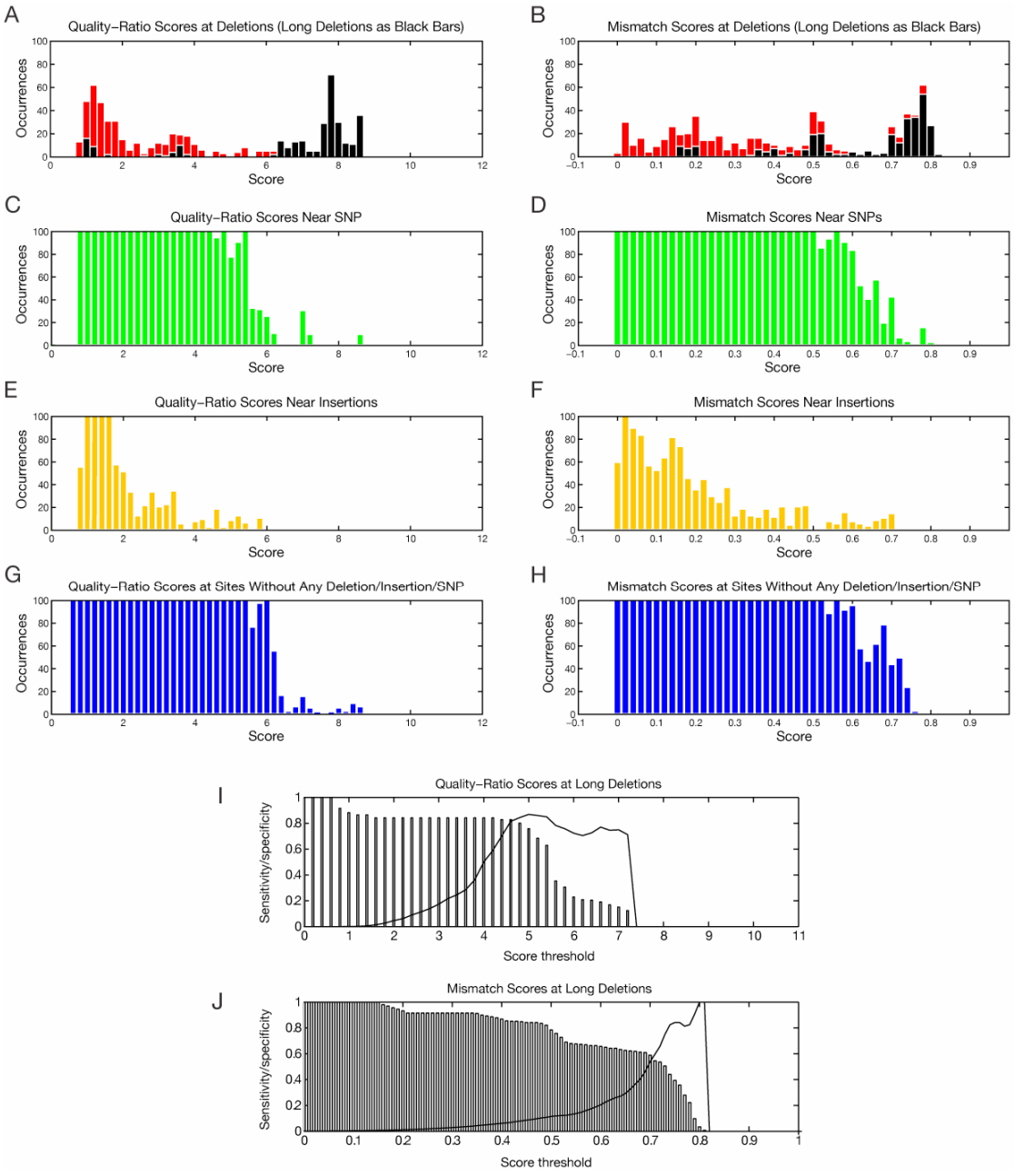

Figure S9. Scores for predicting polymorphic regions partitioned by sequence type and dependency of sensitivity and specificity on score thresholds. Truncated histograms for quality ratio scores (A, C, E, and G) and for mismatch scores (B, D, F, and H) are partitioned by sequence type as labeled. In A and B, red bars denote scores in short deletions and black bars scores for longer deletions (> 25 bp). Scores for 12 bp neighborhoods for SNPs or insertions are shown (C, D, E, and F), as are scores for conserved 25-mers (no polymorphism, G and H). The relationship between sensitivity (bars) and specificity (solid line) as a function of score thresholds (horizontals thin lines) is shown for quality scores (I) and mismatch scores (J). Data are from Br-0, the accession with the largest set of deleted bases in the 2010 dataset.

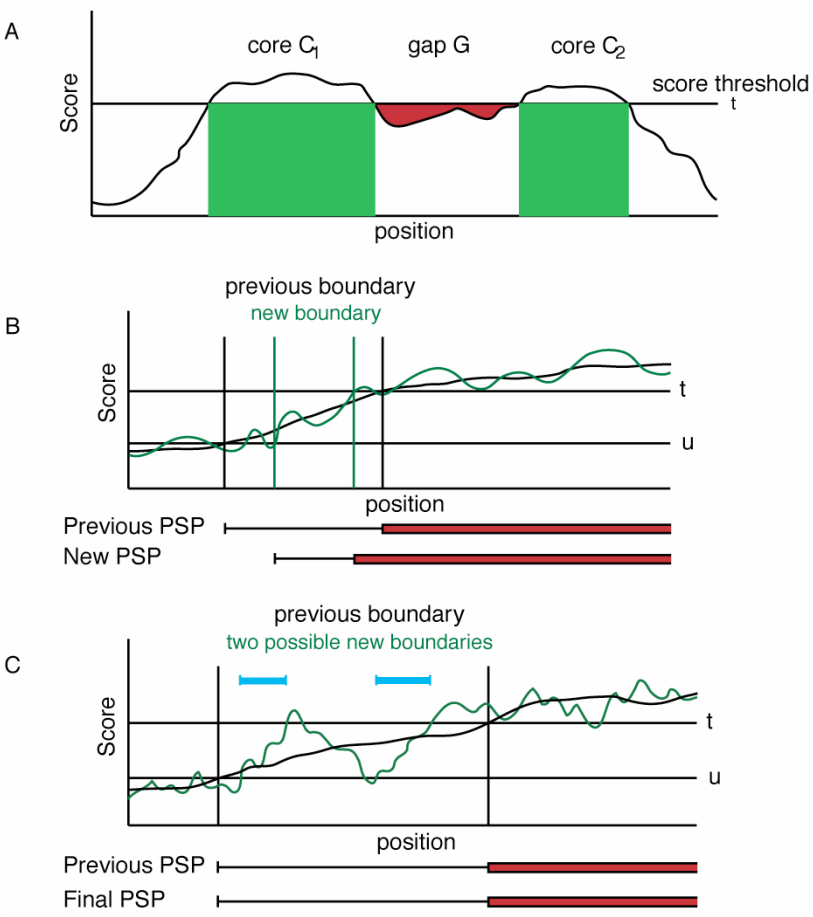

Figure S10. Schematic representation demonstrating core merging and boundary prediction for PRPs. (A) Predicted cores C_1 and C_2 are merged where the sum of the seed lengths (green areas) is greater than twice the length of the intervening region (red area). (B-C) Illustrations of boundary refinement where the black lines indicate scores computed with the original window size and the green lines indicate scores computed with a reduced window size. New boundary regions are computed as shown in panel B, and boundary refinement is terminated in the event of non-intersecting new boundary intervals (shown in blue, panel C). Core predictions are indicated by red bars, with whisker bars denoting boundary regions.

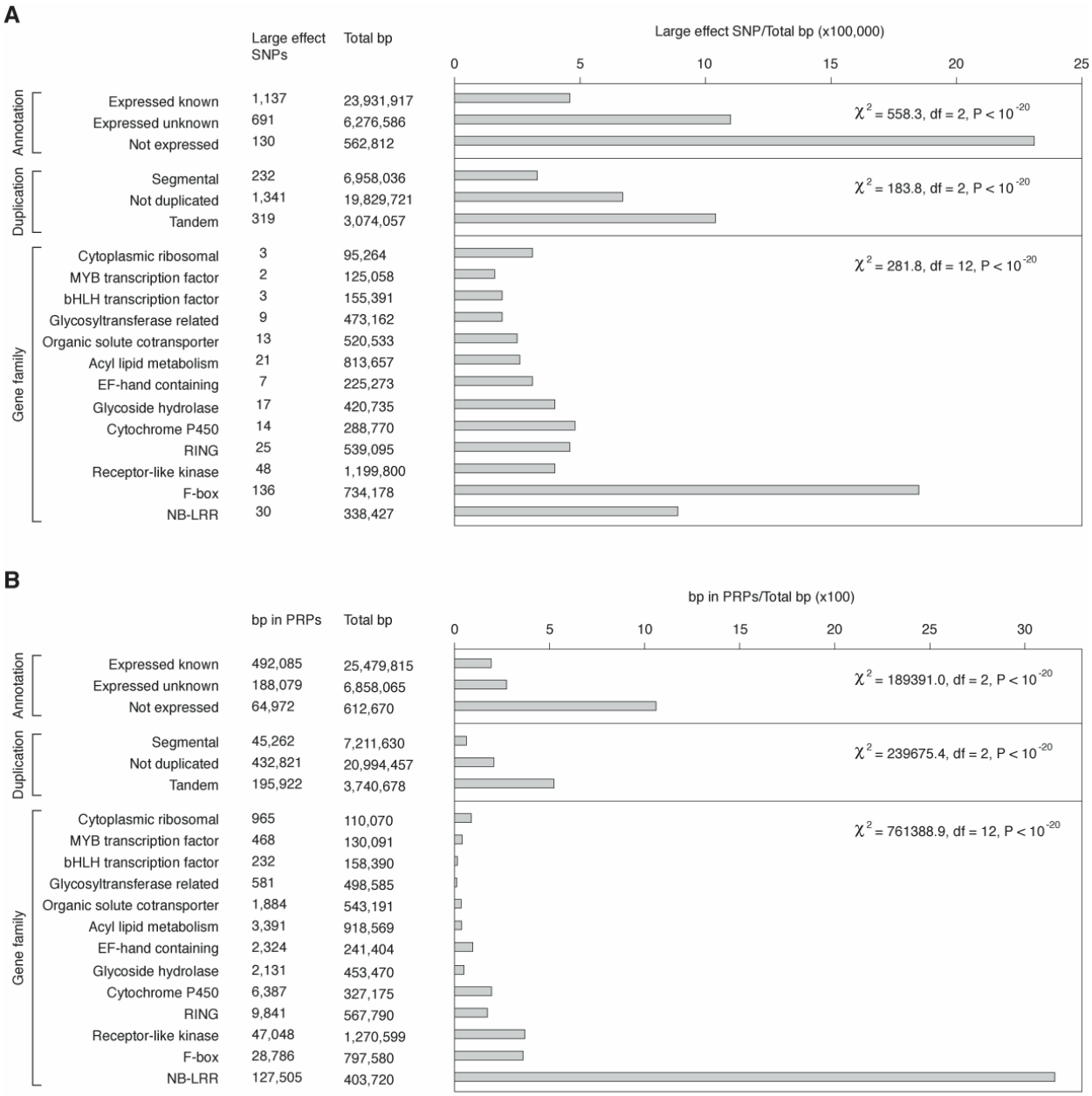

Figure S11. Large-effect SNP and PRP frequency as a function of positions that could be called (genes included in analysis are as for Fig. 3A). (A) Large-effect SNPs normalized by the number of positions for which SNPs could be predicted (i.e., exact and short 25-mer matches excluded). (B) Bases included in PRPs relative to the number of possible bases by category. Differences in total bases between A and B are due to repetitive positions being included in PRPs. In all cases, representation for large-effect SNPs and PRPs differs among categories by Annotation, Duplication, and Gene family groupings (P-values are from χ^2 tests under the null hypothesis that each category is equally represented).

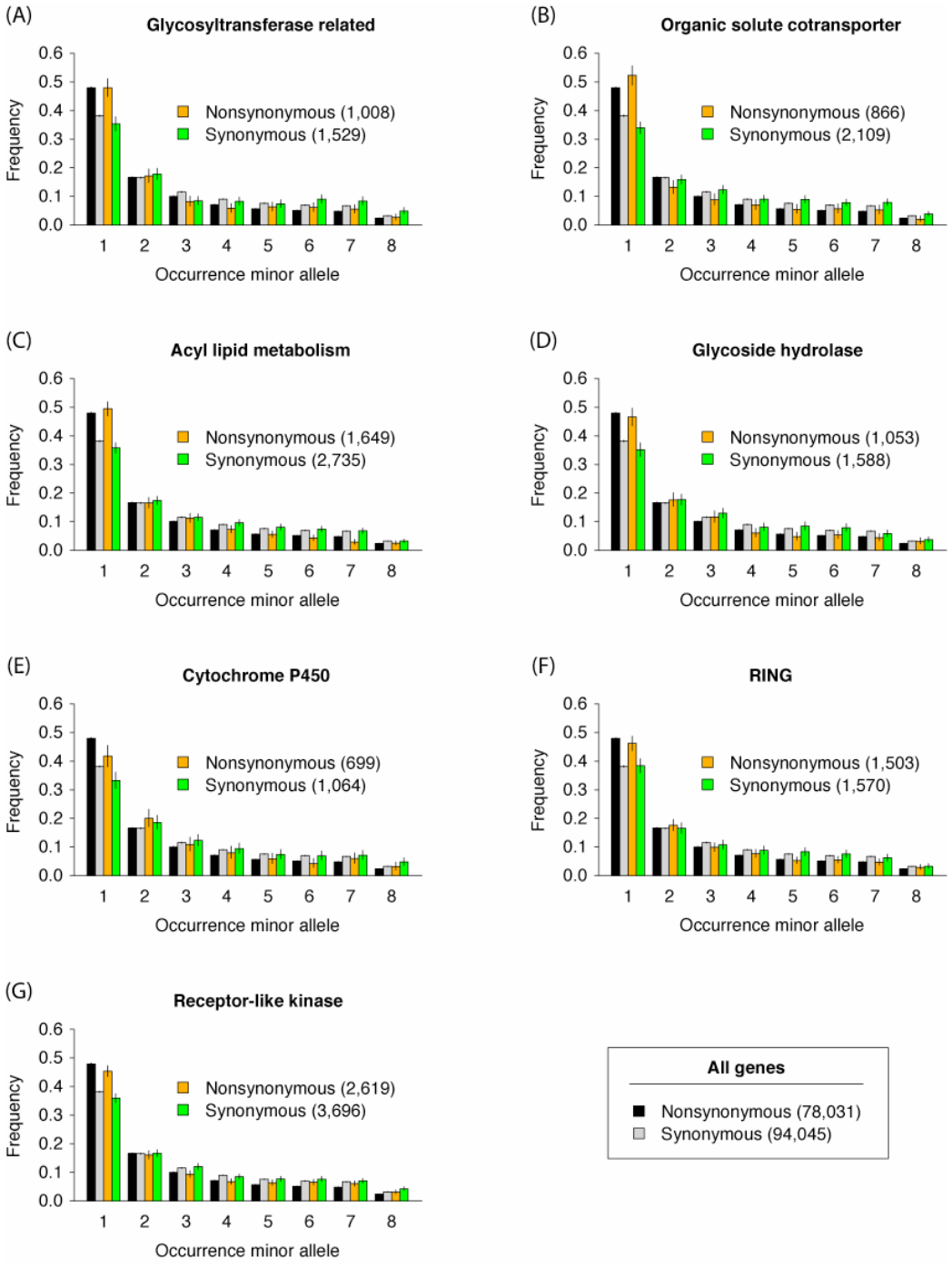

Figure S12. Minor allele frequency by SNP type and gene family where sample size for nonsynonymous and synonymous substitutions by family exceeds 500 (data for NB-LRR and F-box families are given in Fig. 3B). Sample size for all genes at bottom right. Subsampling and error estimates are as for Fig. 3B.

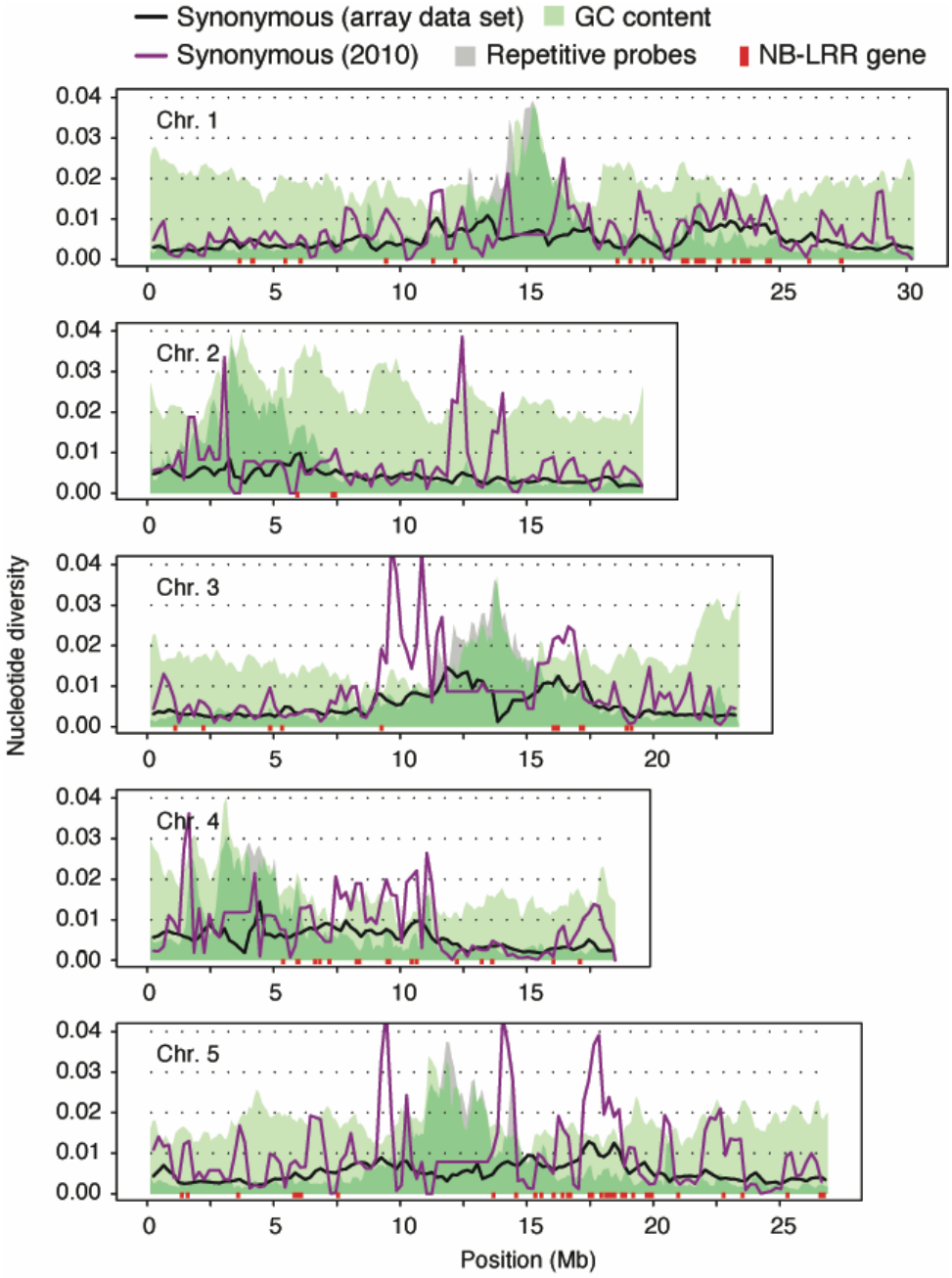

Figure S13. Comparison of genome-wide nucleotide diversity patterns from the array-based and 2010 datasets. Average pairwise nucleotide diversity is plotted for 4-fold degenerate (synonymous) sites for both array-based and 2010 data with sliding windows of 500 kb (counted from all sites) with an offset of 200 kb. GC content in each window calculated from sites called in the Col-0 sample has been rescaled so 35% is at the bottom of each plot and 47.5% is at the top. The content of repetitive probes in each window is rescaled so 100% is at the top of each plot and 0% is at the bottom. Though much sparser, the 2010 dataset supports diversity patterns seen in the array-based data, including increases in diversity flanking centromeres and peaks in diversity in NB-LRR gene clusters. The trend for diversity from the 2010 data to be higher than

that from the array-based data is consistent with the bias against highly polymorphic regions in the array-based pseudochromosomes (since the exact position of potential SNPs in PRPs cannot be determined, they do not factor into diversity estimates). Our estimates of diversity with the array-based data are therefore likely to be underestimates, even for four-fold degenerate sites.

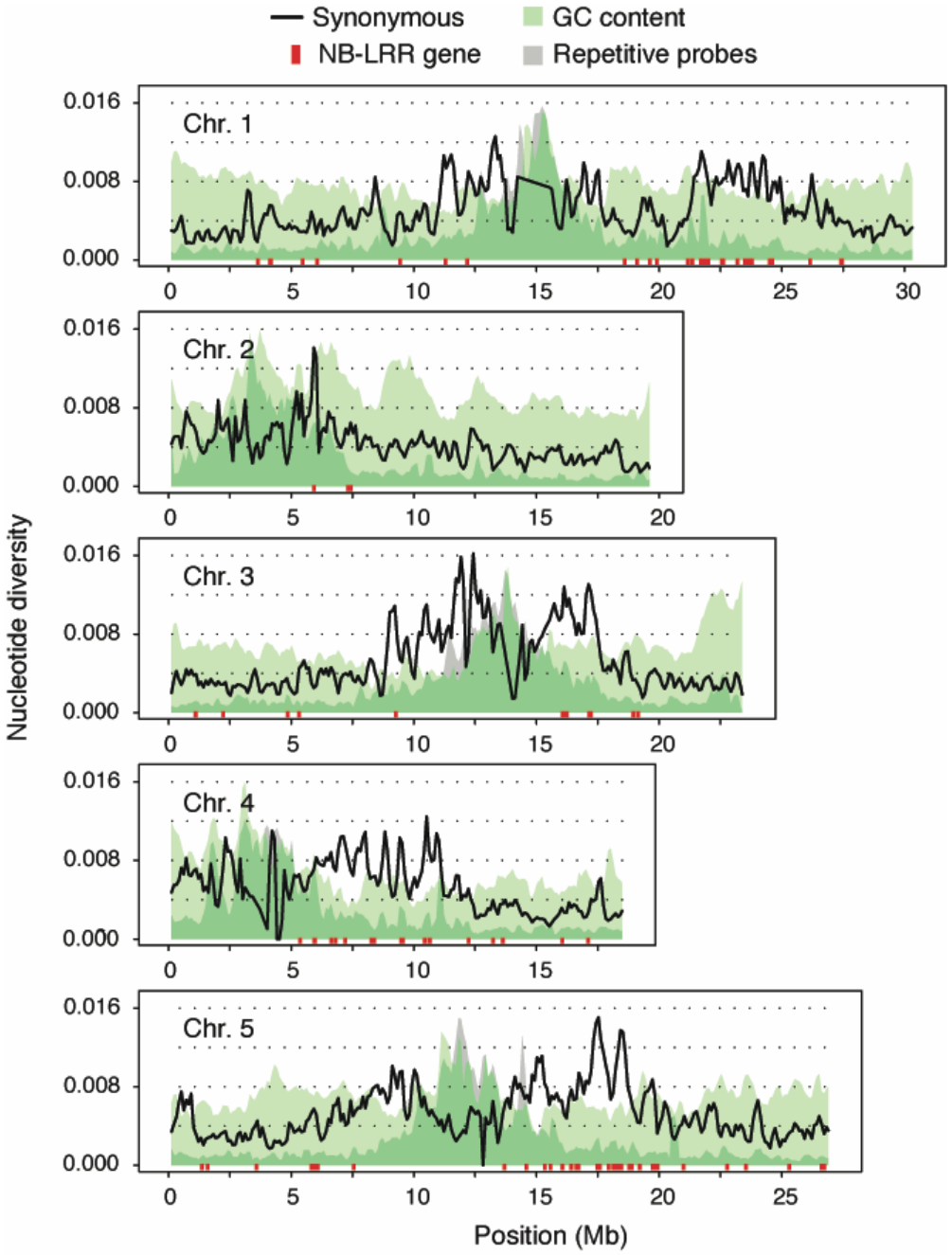

Figure S14. Four-fold degenerate site nucleotide diversity excluding NB-LRR genes. Average pairwise nucleotide diversity is plotted for four-fold degenerate sites along each chromosome with sliding windows of 250 kb (counted from all sites) with an offset of 100 kb. GC content in each window calculated from sites called in the Col-0 sample has been rescaled so 35% is at the bottom of each plot and 47.5% is at the top. The content of repetitive probes in each window has been rescaled so 100% is at the top of each plot and 0% is at the bottom. Diversity remains high in NB-LRR cluster regions, even with these genes removed.

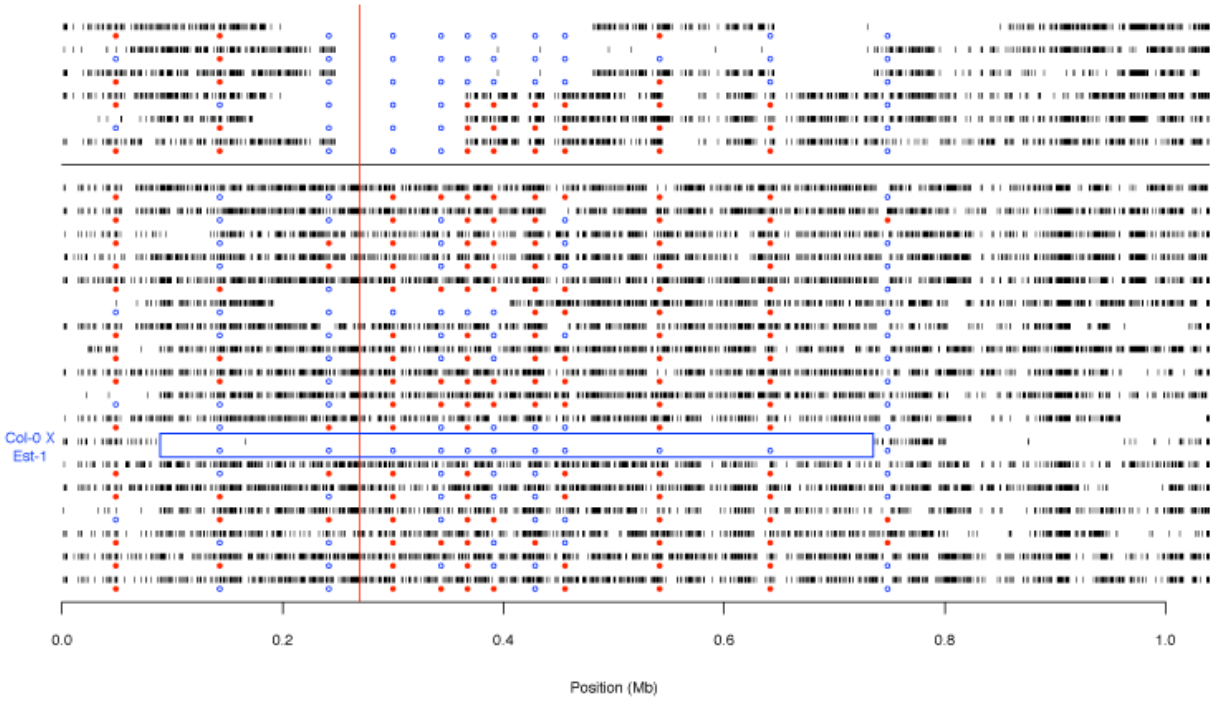

Figure S15. Haplotype sharing in the *FRI* region. Each row represents a comparison between a pair of accessions, with vertical lines indicating the position of mismatches from the MBML2 data and red and open blue circles representing mismatches or matches from the 2010 fragments, respectively. The vertical red line shows the location of *FRI*. The lower 18 rows show comparisons of the Col-0 reference sequence against 18 non-Col-0 accessions (Van-0 excluded). The seventh row from the bottom shows a long region, boxed in blue, of about 600 kb in which Est-1 is almost perfectly identical with Col-0. Est-1 is the only other accession in this sample that carries the Col-0 type deletion in *FRI*. This high similarity was previously apparent from 11 consecutive sequence fragments (open blue circles) which are identical between Col-0 and Est-1 in the 2010 dataset. The top six rows show all pairwise comparisons between the four accessions that carry the Ler-1 type deletion in *FRI*. This set also shows near perfect identity at and around *FRI*, which again was predicted by identity in the 2010 sequence fragments.

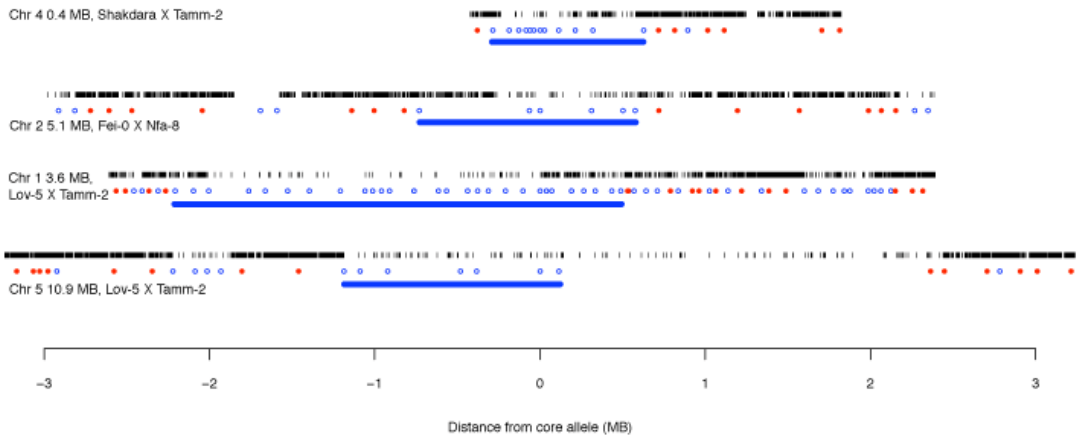

Figure S16. Consistency between previously published regions of extreme haplotype sharing and the current data. We illustrate four low frequency alleles (found in five or six of the 96 accessions) previously identified as located in candidate partial sweep regions (*S36*). In the present data only a pair of accessions share the allele in each case. Identical 2010 sequence fragments are shown as open blue dots, different fragments as closed red dots. Differences in MBML2 SNPs are indicated by vertical lines. The location of each core allele and the accessions that share it are labeled to the left of each row. For these low frequency alleles, we see generally good consistency, as evidenced by the unbroken blocks of identical 2010 fragments (solid blue line) corresponding well to regions of very few mismatches in the MBML2 data. In contrast, higher frequency alleles are more likely to be false positives because unusually high haplotype sharing for these alleles can span relatively few 2010 fragments. Not all high frequency alleles identified as candidates for selection are contradicted in the present data, however, as Toomajian *et al.* (*S36*) did identify as extreme an allele in the chromosome 5 2.8 Mb region with the same accession composition as the second most extreme region of haplotype similarity from the present study (Fig. S19).

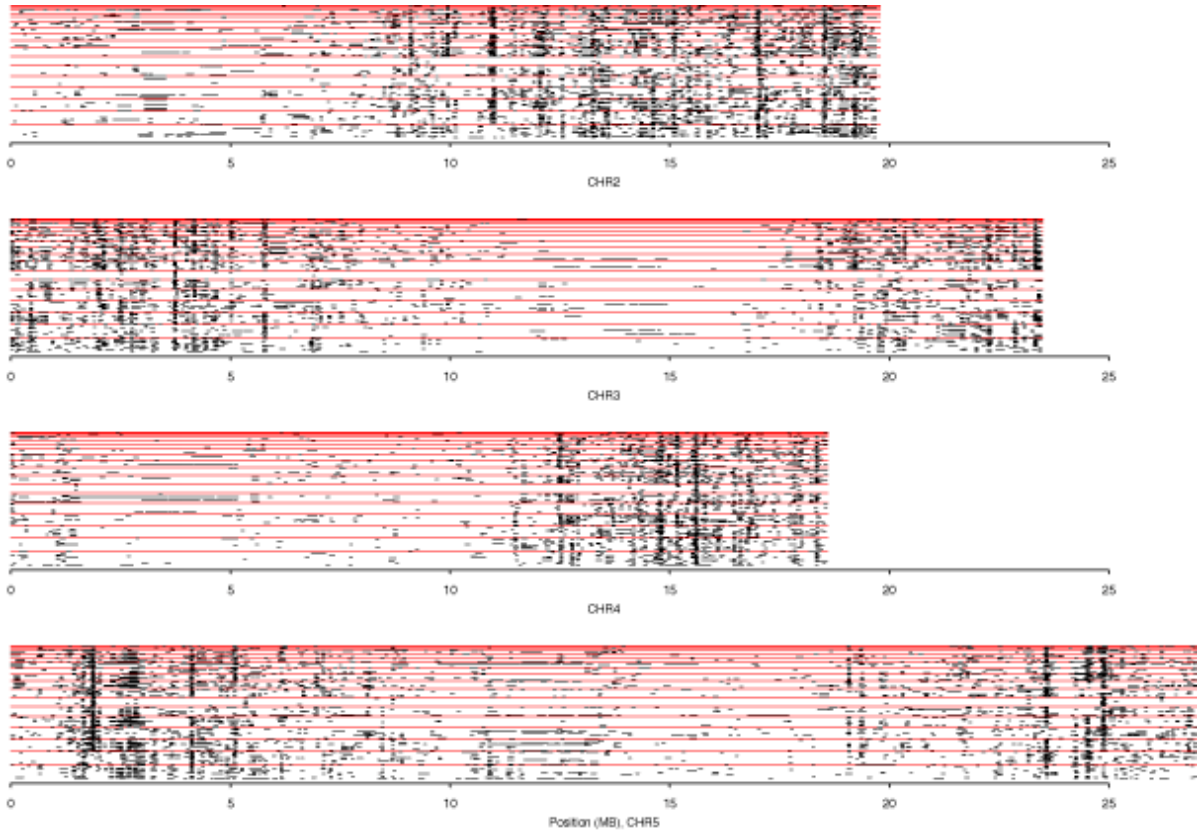

Figure S17. Regions of high pairwise haplotype sharing along chromosomes 2 through 5. Black lines indicate regions of very high similarity between a pair of accessions (rows). Red lines separate comparisons of one accession against the rest. Comparisons are shown only once.

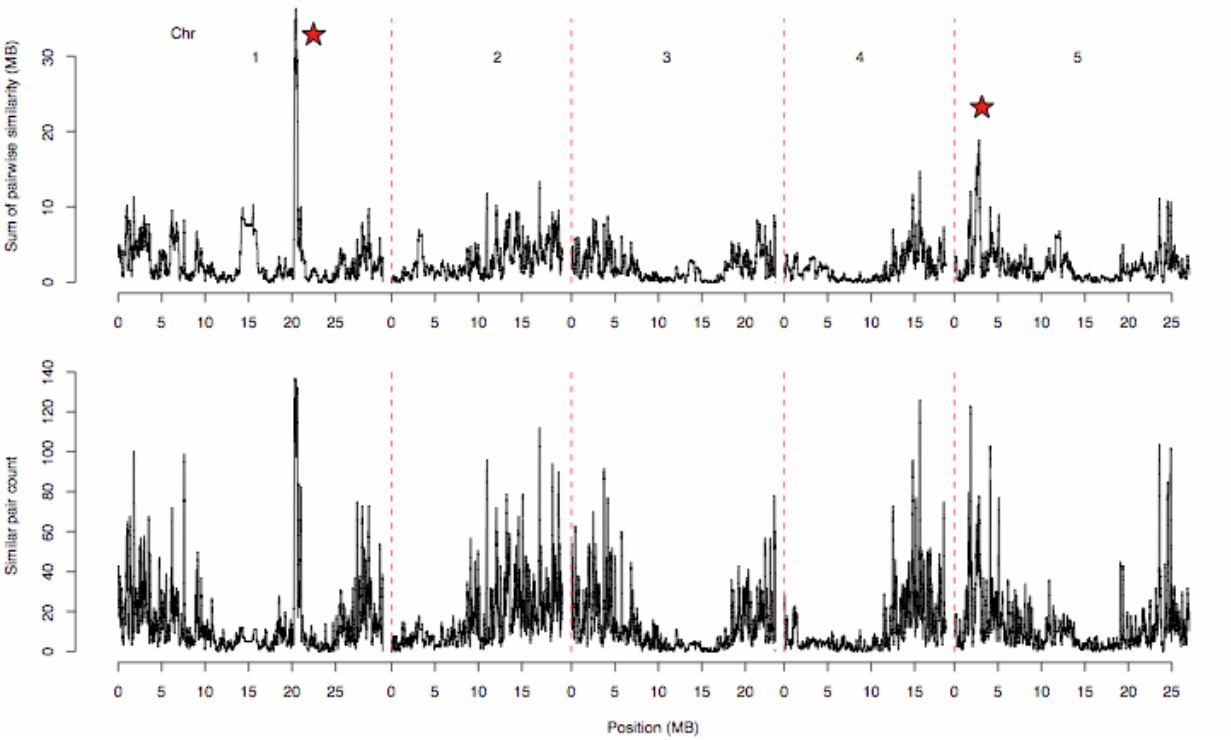

Figure S18. Two simple measures of the extent of high haplotype sharing along all chromosomes. The upper portion plots the total length of runs of high haplotype similarity across all accession pairs in nonoverlapping windows of 10 kb across the genome. The lower portion plots the count (out of a maximum of 171) of accession pairs with high haplotype similarity in nonoverlapping windows of 10 kb across the genome. For the upper portion, the location of the best candidates for partial selective sweeps are indicated by red stars.

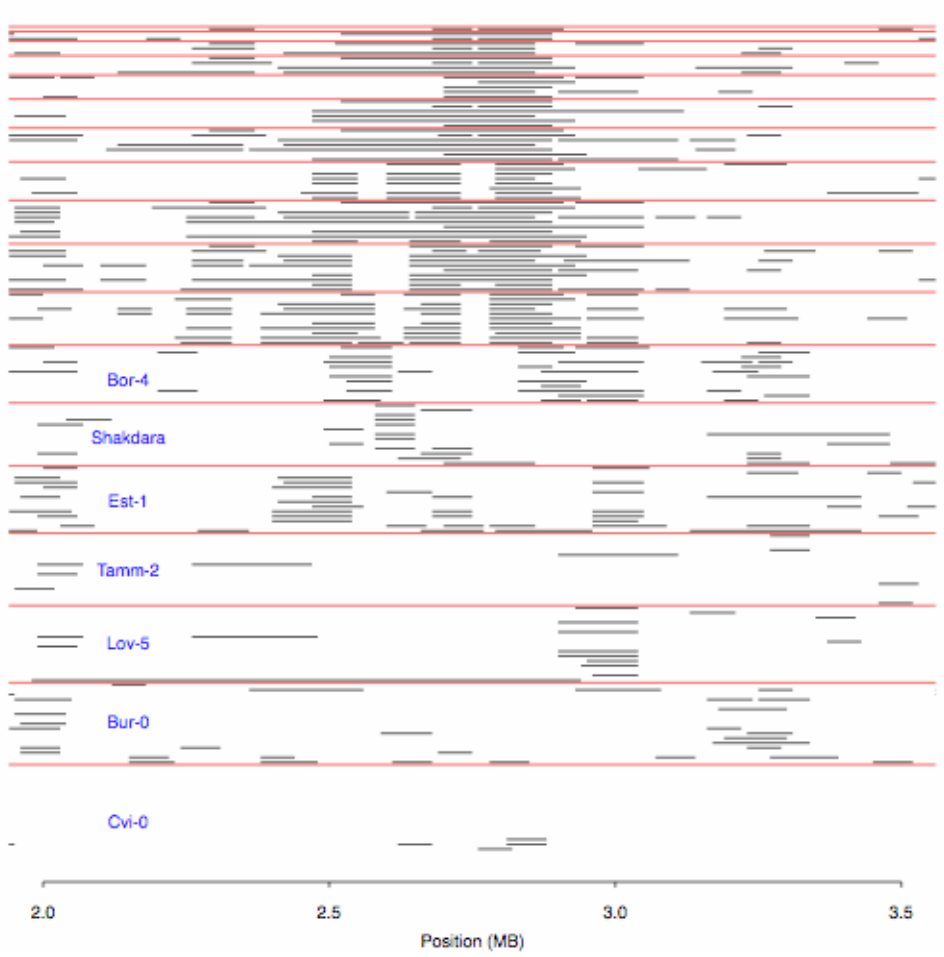

Figure S19. Candidate partial sweep on chromosome 5. The second most striking candidate for a partial sweep is found on chromosome 5, between 2.79 and 2.9 Mb. Black lines indicate regions of very high similarity between a pair of accessions (row). Red lines separate comparisons of one accession against the rest. Comparisons are shown only once. The accession pairs are sorted such that 12 accessions with very high similarity are at the top of the figure. Below these 12, in descending order, Bor-4 is very similar over short stretches to a subset of the 12, but also is similar over longer stretches with Sha and Est-1, which in turn are similar to each other. Tamm-2 and Lov-5 are similar for a very long stretch overlapping this region. Finally, Bur-0 and Cvi-0 are similar to each other as well as to Tamm-2 and Lov-5 over short stretches. The genomic region of highest similarity extends from 2.79 to 2.86 Mb, and includes 19 annotated loci (Table S14).

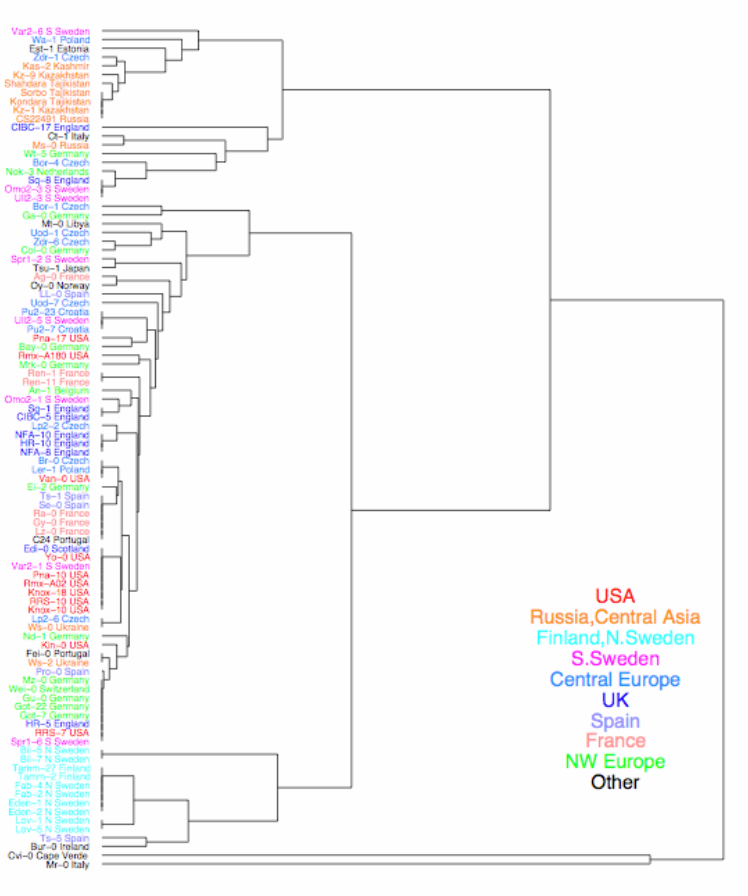

Figure S20. Tree of 11 sequence fragments in the chromosome 5 candidate partial sweep region from 96 accessions. The tree was constructed with hierarchical clustering on the basis of genetic similarity for 11 sequence fragments in the region and shows three major clades and two outliers. One clade includes the 12 similar accessions described in Fig. S19, along with 50 other accessions. A second clade, containing Lov-5 and Tamm-2 as well as Bur-0, is almost exclusively northern Swedish or Finnish. The third clade is predominantly western Asian and Russian and includes Bor-4, Sha, and Est-1. The geographical clustering of the accessions in these clades may represent independent sweeps in each of these geographic areas. While the similarity in the two smaller, or geographically isolated clades might be due to chance, since the similarity was not so extensive here and the accessions involved are similar at many other loci throughout the genome, the major clade is more likely due to a sweep, as this group is typically very heterogeneous across the genome as a whole.

SUPPORTING TABLES

Table S1. Accessions.

Seeds were collected from the material used for hybridization to arrays, and are being distributed by the Arabidopsis Biological Resource Center (ABRC) under the stock numbers indicated.

Accession	Stock number
Bay-0	CS22676
Bor-4	CS22677
Br-0	CS22678
Bur-0	CS22679
C24	CS22680
Col-0	CS22681
Cvi-0	CS22682
Est-1	CS22683
Fei-0	CS22684
Got-7	CS22685
L*er*-1	CS22686
Lov-5	CS22695
Nfa-8	CS22687
Rrs-7	CS22688
Rrs-10	CS22689
Sha (Shakdara)	CS22690
Tamm-2	CS22691
Ts-1	CS22692
Tsu-1	CS22693
Van-0	CS22694

Table S2. Whole-genome repetitive probe set matches for *A. thaliana*.

25-mer match type	Match pairs[a]	Repetitive positions[b]
Exact	333,577,772	12,970,807
Inexact	305,844,001	14,510,324
Short	292,464,314	7,059,270
Union of exact and short[c]	626,042,086	15,537,335
Union of exact, short, and inexact[d]	931,886,087	21,338,048

[a] Pairs of genomic positions with similar probe sequence by match type criteria.
[b] Unique positions tiled on the arrays corresponding to the various repetitive classes.
[c] MB predictions were not generated at these positions.
[d] ML predictions were not generated at these positions.

Table S3. Absolute numbers of MBML2 SNP predictions by target accession and prediction method.

[a] FDRs and recovery evaluated with the 2010 dataset. For "All", FDRs for the MB method adjusted for differences in sequence composition between the 2010 dataset and the genome.

[b] Because reliable Van-0 data are not available from the 2010 dataset, error and recall rates could not be assessed, and ML predictions were not generated.

		SNPs predicted [FDR (%) : Recovery][a]				
Accession	Sequence type	MB	ML	Predicted by MB only	Predicted by both MB and ML	Predicted by ML only
Bay-0	All	97469 [0.7:22.1]	105223 [1.6:22.7]	37798 [1.6:7.2]	59671 [<0.1:15.2]	45552 [4.3:7.7]
	Coding	37919 [0.5:36.8]	39293 [2.0:36.4]	11736 [1.7:11.2]	26183 [<0.1:25.6]	13110 [6.6:10.8]
	UTR+intron	18551 [2.6:15.6]	25994 [2.1:19.0]	5291 [7.5:5.1]	13260 [<0.1:10.5]	12734 [4.7:8.5]
	Intergenic	40999 [<0.1:19.1]	39936 [0.9:18.9]	20771 [<0.1:6.8]	20228 [<0.1:12.2]	19708 [2.6:6.7]
Bor-4	All	94363 [1.9:21.2]	125593 [1.1:25.8]	28040 [7.9:4.4]	66323 [<0.1:17.1]	59270 [2.7:8.9]
	Coding	38211 [1.5:39.7]	44821 [1.7:46.9]	8540 [6.8:8.1]	29671 [<0.1:31.7]	15150 [4.9:15.2]
	UTR+intron	16462 [3.7:12.6]	17018 [1.3:14.9]	7516 [9.4:4.7]	8946 [<0.1:7.9]	8072 [2.7:7.0]
	Intergenic	39690 [1.6:18.9]	63754 [0.6:23.5]	11984 [7.7:3.7]	27706 [<0.1:15.2]	36048 [1.8:8.3]
Br-0	All	88740 [1.2:20.0]	100628 [1.9:20.9]	36837 [3.2:7.6]	51903 [<0.1:12.6]	48725 [4.2:8.6]
	Coding	35844 [0.6:34.8]	32054 [2.0:30.2]	14626 [1.3:15.4]	21218 [<0.1:19.3]	10836 [5.4:10.8]
	UTR+intron	15131 [1.7:12.3]	27254 [1.2:16.6]	3093 [5.9:3.4]	12038 [<0.1:9.0]	15216 [2.7:7.6]
	Intergenic	37765 [1.7:17.9]	41320 [2.2:20.5]	19118 [4.3:6.8]	18647 [<0.1:11.1]	22673 [4.6:9.3]
Bur-0	All	113328 [2.6:21.5]	111401 [2.1:19.9]	48691 [5.0:8.1]	64637 [0.9:13.9]	46764 [4.5:6.3]
	Coding	42427 [0.7:37.8]	43805 [2.2:39.8]	13080 [0.8:9.8]	29347 [0.6:27.9]	14458 [6.6:11.8]
	UTR+intron	21596 [4.6:14.8]	29991 [1.2:16.1]	6092 [11.3:4.8]	15504 [1.0:10.0]	14487 [1.6:6.1]
	Intergenic	49305 [3.5:19.3]	37605 [2.7:15.0]	29519 [5.6:9.5]	19786 [1.4:9.9]	17819 [5.1:5.1]
C24	All	111154 [3.6:21.2]	117308 [1.2:20.2]	43421 [8.9:7.8]	67733 [0.4:13.6]	49575 [2.9:6.9]
	Coding	42932 [2.1:35.7]	41836 [1.4:32.7]	13838 [5.2:12.8]	29094 [0.3:22.9]	12742 [4.6:9.7]
	UTR+intron	20538 [3.0:15.2]	28706 [0.6:15.9]	5588 [6.3:5.7]	14950 [1.0:9.7]	13756 [<0.1:6.2]
	Intergenic	47684 [5.3:19.0]	46766 [1.5:17.7]	23995 [11.6:8.0]	23689 [<0.1:10.9]	23077 [3.8:6.7]
Cvi-0	All	106197 [3.5:16.2]	144355 [1.5:18.7]	34035 [8.6:5.4]	72162 [0.3:10.9]	72193 [3.2:8.0]
	Coding	47055 [1.3:29.9]	50122 [1.3:29.6]	14407 [3.0:10.2]	32648 [0.3:19.8]	17474 [3.1:9.8]
	UTR+intron	18513 [5.3:10.0]	22740 [1.1:12.3]	7452 [10.8:4.0]	11061 [1.1:6.0]	11679 [1.1:6.3]
	Intergenic	40629 [5.3:14.1]	71493 [1.9:17.6]	12176 [13.7:5.0]	28453 [<0.1:9.1]	43040 [3.8:8.4]
Est-1	All	92635 [1.3:20.5]	57233 [1.1:22.8]	56271 [2.6:9.8]	36364 [<0.1:15.1]	20869 [3.0:7.8]
	Coding	36555 [0.9:39.4]	38050 [1.1:40.5]	10642 [2.7:12.5]	25913 [<0.1:26.9]	12137 [3.3:13.7]
	UTR+intron	16638 [0.9:13.4]	14656 [0.8:14.7]	8710 [2.2:5.3]	7928 [<0.1:8.1]	6728 [1.8:6.6]
	Intergenic	39442 [1.9:17.3]	4527 [1.9:8.3]	36919 [2.7:11.9]	2523 [<0.1:5.4]	2004 [5.3:2.9]
Fei-0	All	93129 [1.5:19.4]	116713 [1.7:23.1]	31438 [5.1:5.4]	61691 [<0.1:14.2]	55022 [4.1:9.0]
	Coding	37795 [<0.1:32.9]	47099 [2.0:36.9]	8174 [<0.1:10.2]	29621 [<0.1:22.7]	17478 [5.1:14.2]
	UTR+intron	17020 [5.6:12.0]	22322 [0.7:17.4]	6014 [17.6:3.3]	11006 [<0.1:8.8]	11316 [1.4:8.6]
	Intergenic	38314 [1.1:17.6]	47292 [1.9:19.7]	17250 [3.1:5.9]	21064 [<0.1:11.8]	26228 [4.5:8.0]
Got-7	All	91736 [3.2:19.0]	77946 [1.7:19.1]	47196 [6.4:7.4]	44540 [0.9:11.2]	33406 [2.7:8.0]
	Coding	37908 [1.6:34.9]	27978 [1.8:25.7]	18320 [2.6:17.8]	19588 [0.6:17.0]	8390 [4.2:8.7]
	UTR+intron	16161 [3.3:13.1]	19439 [1.1:19.0]	6322 [11.8:3.3]	9839 [<0.1:9.7]	9600 [2.3:9.3]
	Intergenic	37667 [4.9:16.0]	30529 [2.0:16.0]	22554 [8.0:7.5]	15113 [1.9:8.5]	15416 [2.1:7.5]
L*er*-1	All	92386 [1.9:20.0]	106602 [1.3:20.3]	36606 [4.7:7.2]	55780 [0.2:12.7]	50822 [2.9:7.9]
	Coding	37567 [1.8:35.1]	33283 [2.0:30.0]	14848 [3.7:13.9]	22719 [0.4:21.3]	10564 [5.8:8.8]
	UTR+intron	16448 [2.3:13.0]	21251 [1.2:16.2]	5897 [7.0:4.1]	10551 [<0.1:9.0]	10700 [2.7:7.2]

	Intergenic	38371 [1.9:17.4]	52068 [0.9:18.9]	15861 [4.7:6.8]	22510 [<0.1:10.6]	29558 [2.0:8.3]
Lov-5	All	94938 [2.7:19.9]	83075 [1.0:20.0]	47153 [6.1:8.0]	47785 [0.2:13.1]	35290 [2.6:7.2]
	Coding	39677 [1.7:33.7]	44430 [1.1:37.4]	11181 [5.0:9.5]	28496 [0.3:24.2]	15934 [3.1:13.1]
	UTR+intron	17341 [2.6:14.3]	17572 [2.0:13.9]	8609 [5.8:6.3]	8732 [<0.1:8.0]	8840 [4.7:5.9]
	Intergenic	37920 [3.7:16.9]	2107[<0.1:12.8]	27363 [6.7:9.2]	10557 [<0.1:7.7]	10516 [<0.1:5.1]
Nfa-8	All	95512 [2.3:21.1]	112942 [2.0:20.8]	33707 [6.0:7.0]	61805 [0.2:14.1]	51137 [5.2:6.9]
	Coding	38385 [1.0:35.5]	44421 [1.7:41.2]	9494 [4.4:7.6]	28891 [<0.1:27.9]	15530 [5.1:13.3]
	UTR+intron	17067 [4.6:14.4]	22228 [2.5:18.4]	5869 [10.4:5.0]	11198 [1.2:9.5]	11030 [3.8:8.9]
	Intergenic	40060 [2.6:18.6]	46293 [2.1:15.3]	18344 [5.5:8.7]	21716[<0.1:10.0]’	24577 [5.9:5.3]
Rrs-7	All	93912 [3.8:19.2]	79126 [1.8:20.2]	47680 [8.3:7.7]	46232 [0.4:12.9]	32894 [4.0:7.6]
	Coding	37419 [1.0:37.9]	34751 [1.6:30.5]	13381 [2.2:16.2]	24038 [<0.1:21.6]	10713 [5.2:8.9]
	UTR+intron	16146 [4.5:11.9]	28766 [1.8:18.3]	3044 [19.2:2.4]	13102 [<0.1:9.5]	15664 [3.7:8.8]
	Intergenic	40347 [6.1:17.2]	15609 [2.3:13.8]	31255 [9.8:8.8]	9092 [1.9:8.3]	6517 [2.9:5.5]
Rrs-10	All	97455 [2.5:23.1]	102635 [2.1:22.5]	37983 [6.9:8.3]	59472 [0.3:15.2]	43163 [5.9:7.0]
	Coding	38849 [0.3:38.3]	44086 [1.7:42.3]	9431 [1.1:9.7]	29418 [<0.1:28.6]	14668 [5.1:13.7]
	UTR+intron	17822 [3.4:14.4]	27691 [1.5:17.2]	4177 [6.5:5.5]	13645 [1.4:8.9]	14046 [1.5:8.3]
	Intergenic	40784 [4.3:21.9]	30858 [3.4:17.0]	24375 [9.3:9.7]	16409 [<0.1:12.2]	14449 [11.1:4.7]
Sha	All	95660 [2.8:16.9]	122145 [2.0:18.4]	30248 [7.8:5.5]	65412 [<0.1:11.5]	56733 [4.8:7.0]
	Coding	40184 [1.0:37.5]	50714 [2.0:41.5]	8209 [3.5:10.3]	31975 [<0.1:27.4]	18739 [5.6:14.1]
	UTR+intron	17033 [4.8:10.9]	23941 [2.0:13.8]	5388 [12.2:4.0]	11645 [<0.1:6.9]	12296 [3.9:6.9]
	Intergenic	38443 [3.8:13.5]	47490 [2.0:13.3]	16651 [8.5:5.8]	21792 [<0.1:7.7]	25698 [4.5:5.7]
Tamm-2	All	97447 [4.1:21.0]	108826 [1.0:25.3]	37237 [12.2:6.5]	60210 [0.2:16.3]	48616 [2.2:9.3]
	Coding	40288 [1.3:37.8]	55623 [1.4:45.7]	6223 [5.6:8.1]	34065 [<0.1:29.8]	21558 [4.0:15.9]
	UTR+intron	17413 [3.4:14.2]	29288 [1.0:19.3]	3890 [11.1:3.2]	13523 [0.9:11.0]	15765 [1.2:8.3]
	Intergenic	39746 [7.3:18.4]	23915 [<0.1:15.7]	27124 [13.8:9.0]	12622 [<0.1:9.6]	11293 [<0.1:6.1]
Ts-1	All	93766 [2.2:19.0]	120650 [1.1:21.9]	29960 [6.9:5.5]	63806 [<0.1:13.7]	56844 [2.6:8.5]
	Coding	38333 [1.5:34.8]	48754 [1.5:40.6]	7878 [5.3:9.6]	30455 [<0.1:25.3]	18299 [4.0:15.3]
	UTR+intron	16329 [2.7:10.7]	23619 [0.7:13.5]	5171 [7.9:3.4]	11158 [<0.1:7.3]	12461 [1.6:6.1]
	Intergenic	39104 [2.7:17.8]	48277 [0.8:19.4]	16911 [7.3:6.2]	22193 [<0.1:11.6]	26084 [2.0:7.8]
Tsu-1	All	96107 [2.9:20.5]	80256 [1.8:21.6]	47339 [8.0:7.6]	48768 [<0.1:14.0]	31488 [4.7:7.6]
	Coding	38466 [0.8:34.0]	40652 [1.7:32.4]	10812 [2.4:11.3]	27654 [<0.1:22.6]	12998 [5.5:9.7]
	UTR+intron	17241 [3.4:12.0]	22922 [2.5:16.2]	5590 [10.0:3.8]	11651 [<0.1:8.3]	11271 [5.1:7.9]
	Intergenic	40400 [4.7:20.3]	16682 [1.0:17.0]	30937 [9.5:9.4]	9463 [<0.1:10.9]	7219 [2.6:6.1]
Van-0 [b]	All	93532	NA	NA	NA	NA
	Coding	38224	NA	NA	NA	NA
	UTR+intron	16157	NA	NA	NA	NA
	Intergenic	39151	NA	NA	NA	NA

Table S4. Effect of filters on set 2010 composition.

	Total positions	Total polymorphic positions	Mean no. positions per accession (rounded)	Mean no. polymorphic positions per accession (rounded)
Without filters	674,315	12,967	610,000	2,700
After filter 1	70,968	8,615	7,500	1,900
After filter 2	11,191	6,579	3,200	1,400

Table S5. List of properties that constitute the input vector $\boldsymbol{x}^{(1)}$ at a given position p. If not specified otherwise $\Delta p \in \{-4,...,4\}, \tau \in \{t,col\}, s \in \{+,-\}, \sigma \in \Sigma, \Sigma = \{'A','C','G','T'\}$.

Symbol	Formula	Description	Size
I_{max}	$I^p_{\max}(\Delta p,\tau,s) = \max_{\sigma\in\Sigma} I^s_\tau(p+\Delta p,\sigma)$	maximal intensities for target and reference accession on forward and reverse strand taken in window of length 9	36
I_{sec}	$I^p_{\text{sec}}(\Delta p,\tau,s) = \operatorname{mean}_{\sigma\neq\sigma_{\max}} I^s_\tau(p+\Delta p,\sigma)$ *where* $\sigma_{\max} = \arg\max_{\sigma\in\Sigma} I^s_\tau(p+\Delta p,\sigma)$	average of non-maximal intensities for target and reference accession on forward and reverse strand taken in window of length 9	36
Q_1	$Q^p_1(\Delta p,\tau,s) = I^p_{\max}(\Delta p,\tau,s) / I^p_{\max}(0,\tau,s)$ *where* $\Delta p \in \{-4,...,-1,1,...,4\}$	quotients of maximum intensities at neighboring positions p+Δp and the considered position p, for target and reference accession on forward and reverse strand taken in window of length 9	32
Q_2	$Q^p_2(\Delta p,s) = I^p_{\max}(\Delta p,t,s) / I^p_{\max}(\Delta p,col,s$	quotients between the maximum intensities of the target and the reference accession on forward and reverse strand taken in window of length 9	18
k	$k^p(\Delta p,\sigma) = \left[k^p_{type}(\Delta p,\sigma), k^p_{dom,type}(\Delta p)\right]$ *where* $type \in \{exact,inexact,short\}$,	number of repeated 25mers for each position in the window, (exact, inexact and short 25mers are taken with respect to each possible base, dominating 25mers comprise all dominating 25mers)	135
M	$M^p(\Delta p,\tau,s) = \delta\{B^s_\tau(p+\Delta p), seq(p)\}$ *where* $\delta\{i,j\} = \begin{cases} 1 & if \quad i=j \\ 0 & otherwise \end{cases}$	mismatches between maximum base call and reference sequence for target and reference accession on forward and reverse strand taken in window of length 9	36
seq	$seq^p(\sigma) = \delta\{seq(p),\sigma\}$	binary vector denoting the reference base at the considered position	4
f	$f^p(\sigma) = \sum_{\Delta=-13}^{\Delta=13} \delta\{seq(p+\Delta),\sigma\}$	frequency of each letter of the alphabet □ within the 25mer	4
S	$S^p = -\sum_{\sigma\in\Sigma} f_p(\sigma)\cdot\log(f_p(\sigma))$	sequence entropy of the corresponding probe	1
$\boldsymbol{x}^{(1)}$	[I_{max}, I_{sec}, Q_1 ,Q_2, k, M, seq, f , S]		302

Table S6. Input vector $\boldsymbol{x}^{(2)}$ at position *p* for layer 2 SVMs.

Symbol	Formula	Description	Size
$\boldsymbol{x}^{(1)}$		As described in Table S5	302
b	$b^p(a) = \delta\{p \in p_a\}$ *where* $\delta\{true\} = 1 \quad \delta\{false\} = 0$	binary vector describing whether position *p* passed filter 1 for accession *a*	18
c	$c^p(a)$ (see Section 6)	transformed output values of SVM 1 at position *p* of accession *a*	18
$\boldsymbol{x}^{(2)}$	$[\boldsymbol{x}^{(1)},b,c]$		338

Table S7. Number and bases included in PRPs by accession.

Accession	***n***	**Cores only**	**Cores + Boundaries**
Bay-0	713	1,053,867	1,198,126
Bor-4	601	725,557	937,013
Br-0	758	1,065,389	1,294,414
Bur-0	663	847,274	1,122,014
C24	770	884,482	1,004,570
Cvi-0	1019	1,413,710	1,555,356
Est-1	320	406,850	498,031
Fei-0	674	942,816	1,088,730
Got-7	610	799,245	1,066,782
L*er*-1	849	1,192,448	1,452,623
Lov-5	737	1,118,765	1,088,989
Nfa-8	801	1,143,879	1,414,009
Rrs-7	696	962,922	1,502,504
Rrs-10	605	818,190	995,484
Sha	774	1,228,239	1,508,522
Tamm-2	770	1,142,890	1,299,966
Ts-1	763	1,073,443	1,254,375
Tsu-1	628	823,264	1,044,783
Van-0	719	1,040,583	1,316,483

Table S8. Genome-wide percentages of bases called as reference sequence by accession and sequence type.

Accession	Coding	UTR	Intron	Inter-genic	Pseudo-gene	Trans-poson	All
Bay-0	85.9	66.3	64.7	53.5	55.0	46.1	65.9
Bor-4	87.0	62.5	59.5	49.3	59.0	54.2	63.6
Br-0	82.3	55.1	53.0	43.5	50.1	43.6	57.7
Bur-0	86.9	71.2	70.3	59.8	58.9	51.4	70.3
C24	88.1	66.8	65.1	52.6	53.4	48.2	66.4
Col-0	92.3	70.9	69.9	61.2	75.5	81.4	73.5
Cvi-0	80.3	52.6	49.8	39.7	45.6	41.1	54.7
Est-1	89.1	66.2	63.8	53.2	61.7	55.4	66.9
Fei-0	84.3	60.7	58.4	47.8	51.7	46.9	61.6
Got-7	83.6	58.9	55.5	46.5	54.1	50.7	60.3
L*er*-1	82.7	58.8	56.2	45.8	45.6	39.6	59.4
Lov-5	80.4	56.1	53.4	43.5	43.8	38.5	57.1
Nfa-8	84.7	59.9	57.4	46.4	50.0	44.9	60.8
Rrs-7	85.1	60.6	57.6	48.4	52.7	49.5	62.0
Rrs-10	88.8	67.0	64.6	54.1	60.7	53.5	67.4
Sha	81.6	55.8	53.3	42.8	43.8	38.7	57.1
Tamm-2	83.5	58.3	55.9	45.6	47.6	42.3	59.6
Ts-1	83.3	57.8	55.1	45.0	47.0	44.9	59.2
Tsu-1	88.2	64.5	61.8	51.3	58.9	54.7	65.3
Van-0	84.9	59.3	56.4	46.5	52.9	49.3	60.9

Positions with exact, short, and inexact 25-mer matches not included in calculating percentages.

Table S9. Predicted large-effect SNPs and empirically determined FDRs.

			Validation by dideoxy sequencing			
Effect of SNP		***n***	**Attempted**	**True**	**False**	**FDR**
Premature stop		1,227	612	413	38	0.08
Stop codon converted to coding		198	89	59	4	0.06
Loss of initiation methionine		156	56	37	2	0.05
Splice donor:	Knockout	145	64	44	3	0.06
	GT to GC [a]	77	27	22	0	ND
	GC to GT [a]	14	10	7	0	ND
Splice acceptor		290	102	68	4	0.06
All		2,107	960	650	51	0.07

[a] Consensus-to-nonconsensus splice changes (or vice versa) are here reported here, but were not considered as "large-effect SNPs" for other analyses.

Table S10. Status for validation by dideoxy sequencing of large-effect SNPs.

Notes:
[a] Chromosome
[b] PreStop: premature stop codon
RevStop: stop in Col-0 not a stop in another accession
Met: initiation methionine changed to another amino acid
SA: nonfunctional splice acceptor change
SD: nonfunctional splice donor change
SD(non): consensus splice donor in Col-0 changed to nonconsensus (GT to GC)
SD(con): nonconsensus splice donor in Col-0 changed to consensus splice donor (GC to GT)
Note that, while consensus to nonconsensus splice donor changes are reported, they were not considered as "large-effect SNPs" for most analyses (see Section 10).
[c] "False" indicates that the reference base or a third base were present.

Gene	Chr. [a]	Position	Acces-sion	Predic-tion	Effect [b]	Valida-tion [c]	Primers used for validation (forward, reverse)
AT1G01180	1	77140	Nfa-8	G->A	PreStop	True	ctgttcacatttcggttaagg, gctcttggcaataagatgagc
AT1G01440	1	159935	Tamm-2	T->A	RevStop	False	tatttcccaacaggcaacagg, gagttcttgatggagactctgg
AT1G01450	1	165397	Cvi-0	G->A	PreStop	True	ccacatacctctttgatgtgc, atgtttacctggaagtttggg
AT1G01590	1	214406	Fei-0	G->T	PreStop	True	ttcttggaagttcatctctgg, aaagagtgagcagtcgttagc
AT1G02300	1	454975	Cvi-0	A->T	PreStop	True	atcaaaacactatggtgtcgg, aagaacttacatcaccccagc
AT1G02620	1	557507	Rrs-7	G->T	PreStop	True	gtcgtctctcgaagctaaagg, gaagaatgaaggcttctctgg
AT1G02670	1	576108	Bay-0	T->A	PreStop	True	cagaatcttgagttttgtggg, acacgtgtcgatttcttcg
AT1G02990	1	683022	Bay-0	C->T	SD	True	ctgcttgcaactcatatctgc, agatccaaggatatttacggc
AT1G03300	1	813018	Bur-0	G->C	PreStop	True	tgtaatcagcatcaaccatcg, aaggagtgatcttatctgggc
AT1G03300	1	812120	Van-0	T->A	PreStop	True	ctctcttccttttgtgctcc, aggctaccaaggataagttgg
AT1G03420	1	847403	Lov-5	G->A	PreStop	True	acaaactggatctgagatggc, gtcaccggaaaaagagaacc
AT1G04710	1	1324306	Ts-1	G->T	SA	True	aatcctactgtgtgttcaggc, ggacatcacagagctcatcc
AT1G04790	1	1345987	Sha	G->T	PreStop	True	gaaatccatctccactgatcc, tcttttgcctctagctcttcc
AT1G05220	1	1512239	Bur-0	G->T	PreStop	True	ctatttcctggaatctacccg, acaaagcggtacaatctagcc
AT1G05830	1	1759022	Est-1	G->A	SD	True	acttcaaaacaggatcgttgg, aatttgacgtttgtcgatgc
AT1G06840	1	2103476	Cvi-0	C->A	Met	True	gaggaagaagaagagcagagg, atcaaggtgggataaaaagg
AT1G07025	1	2157705	Fei-0	C->T	PreStop	True	aggaaatctcaggtgacaagc, atggtatatgaggcatggagc
AT1G07280	1	2239701	Cvi-0	G->T	PreStop	False	cgattcttcaatggttaagacg, tgtgcgtaattagcaagaagg
AT1G07330	1	2253513	Cvi-0	C->T	PreStop	True	tgactctgatgaacctgaagc, cttctctagcttctcacttgcc

AT1G08300	1	2615115	Rrs-10	C->T	PreStop	True	ggtacgtttcaccttaaaccc, aacatcttgcatccatatccc
AT1G09140	1	2942889	Rrs-7	T->C	RevStop	True	cgagacagagtttccggc, atcgaattctccagtttcacc
AT1G09320	1	3012230	Rrs-7	T->C	Met	True	tccttatctgaggaagatggg, gctacaacctctttttagccc
AT1G09400	1	3033678	Cvi-0	T->G	RevStop	True	acattctacacattgcttccg, gtcagtgagattttcgagtgc
AT1G09950	1	3240916	Sha	T->A	PreStop	True	ttaaggaagaaacgagaagcc, gagatttcgccctcgtgc
AT1G10210	1	3349671	Br-0	A->T	PreStop	True	gattccgatcgtttatgttcc, acattctcatgtcgaagatgg
AT1G10540	1	3476423	L*er*-1	A->G	SD(non)	True	tgtatgtccctgtagactgttcc, ttgatgtgtgcagtacattcc
AT1G10570	1	3490724	Bay-0	C->A	SD	True	taatgtcaaaaggtgagaccc, caaagatggaagaaaaacgc
AT1G10660	1	3532917	C24	A->G	SA	True	cctctcaatcttcgaatctcc, gcgttgtcttcttctcttgc
AT1G10680	1	3541086	Tamm-2	T->A	PreStop	True	gtctccgtgattggtagttcc, agaggaattcagctatctggc
AT1G10880	1	3624082	Bay-0	C->A	PreStop	True	gctctggacatagctactagaatcc, acaggcaatggagttaaaagg
AT1G11160	1	3738010	Bur-0	T->A	PreStop	False	tatgatgagccttctgtagcg, agtacctgaggcatgttatcg
AT1G11180	1	3747407	Bay-0	G->A	PreStop	True	cttttctggaatatcatcgcc, acagtgccactaaaaccagc
AT1G11925	1	4026309	Br-0	C->T	SD	True	gaacaagtactagaccgttatgtaagc, aaatatggttgggggatagc
AT1G12350	1	4200106	Rrs-10	C->T	SD(con)	True	tttatctcaatgcttgtggg, gagaatggagaaggagagagc
AT1G12660	1	4311015	Bay-0	C->T	PreStop	True	ctgatttgtggatgaatctgg, acttactcggattttggatgg
AT1G12700	1	4325389	Fei-0	C->T	PreStop	True	ttctgcatacaatacccatcc, gagatacttctttggccttgg
AT1G12700	1	4325709	Fei-0	T->A	PreStop	True	cttgaaaaggctaattgcagc, tcgtgttagatttctgcaagc
AT1G13430	1	4607001	Fei-0	A->T	PreStop	True	gttcaacgatcaaaacactcg, aatagctctgtgcatggtcc
AT1G13490	1	4624903	Rrs-7	C->A	PreStop	False	cacacataacacacaaagaagc, tttgaggttgtagttgatgtgg
AT1G13510	1	4630397	Rrs-7	G->T	PreStop	True	atctagctgttttggtttggc, cgaggtgtttgtcagagtacc
AT1G13770	1	4723657	Cvi-0	G->A	PreStop	False	agaatttctctcactgtttgcc, gaagctccaacaccatttagc
AT1G13780	1	4725412	Tsu-1	A->T	PreStop	True	ctttgaagctttcgtattcgg, ttctctttggtaagtcctccg
AT1G15165	1	5218613	Cvi-0	C->T	PreStop	True	caaataaacacgagggtatgc, tgcactttattacaaggtgtgg
AT1G15590	1	5368416	Sha	T->A	RevStop	True	tctgtaagggaaggtataggagg, cgtatctttcagtttcccacc
AT1G15680	1	5394482	Nfa-8	A->T	PreStop	True	cttctattgaggctttggagg, ttcatgatgcgtacatatccc
AT1G16025	1	5501643	Bur-0	G->T	SD	True	aaccagtaaagagaacggagg, cacatctcaaatccacaaagc
AT1G16260	1	5559780	C24	G->T	PreStop	True	cgtttttgtccttgatttgc, agaagaggtccttgcagtagc
AT1G16260	1	5561781	Tsu-1	G->A	PreStop	True	acatcccacgatttgtaacc, gagagagagagagcaaatggg
AT1G17120	1	5851958	Ts-1	T->C	Met	True	actttaggtgggtccctcg, caacagaaccaaagctaagcc

AT1G17450	1	5988638	Rrs-7	A->G	RevStop	True	aaaaccttccaaaactcacc, tgtctctatatacggtggttgc
AT1G17890	1	6155588	Est-1	A->G	Met	False	gcggatttctctaacataaacg, tggaaaagtaagcgaaatgg
AT1G18200	1	6264406	Bor-4	C->T	SA	True	taacgtaacctttctgctcc, tgaagaagcgatagtgaatgc
AT1G18410	1	6342093	Bur-0	C->T	SD	True	tttttcttcgtctctgtagcg, ctgagagctgtgagtctgagg
AT1G19060	1	6582314	Cvi-0	T->A	PreStop	True	tgatgtcttctagatacgcgg, gtactctgtgcaattcaaaagg
AT1G19090	1	6591086	Br-0	G->A	SD	False	cagcttcagtgattaaaaccg, gctccaagagaagtaagaatcg
AT1G19490	1	6752771	Est-1	A->C	SD	False	gaagtagggaattcaagtggc, agaatgagaatttgaggaggg
AT1G20320	1	7034392	Bor-4	C->T	PreStop	False	tctgcgtaagattttgactcc, cttgccttaacgagagaaagc
AT1G20370	1	7051889	L*er*-1	C->A	PreStop	True	ggacacactgaaacaaatgtcc, ggtttaaaggcagtcatgagg
AT1G20400	1	7074743	Tsu-1	G->A	PreStop	True	tcatctccttaacatcgtccc, agttttctcaggattcttcgc
AT1G20730	1	7198419	Rrs-10	C->A	PreStop	True	aaacgacgggagtatcatagc, ataatctcctgctcatctccc
AT1G20750	1	7204510	Van-0	C->A	PreStop	True	catgttcaaaggtgactctgc, gttaaaagggaagctgaatgc
AT1G21060	1	7371789	Bur-0	T->C	Met	False	attttctcgctcattttccc, ttgctcagaagaaactaaccg
AT1G21170	1	7418334	Bur-0	T->C	SD(non)	True	tcttagagaattgatgcaccg, aagaattttgtccaggagtgg
AT1G21312	1	7463716	Got-7	C->A	PreStop	True	gggatacttcccttcttgagc, tttatcacaggatgaggatcg
AT1G21860	1	7671207	Rrs-10	G->T	PreStop	True	actcatatcgctcatctgtgg, aaatgaccgttataccaaccg
AT1G21990	1	7742095	Bur-0	T->C	Met	True	aagcctaaagaacagcgacc, tttttcctcaacctcatctgg
AT1G22010	1	7749757	L*er*-1	C->A	PreStop	True	gtcgaaaaagtatccatcaagc, atgttccagttaggctcttcg
AT1G22080	1	7794188	Sha	T->C	SA	True	aaactgtgggatctctttgg, aacaagcaagaagaacagcc
AT1G22290	1	7877497	Bur-0	T->A	PreStop	True	gtaacaacaaccaacattgcc, ggggtacaagaatgtgattagc
AT1G22570	1	7978280	Cvi-0	G->A	PreStop	True	caaattcagctaaaatcccg, aagattacgttagcgattccg
AT1G22980	1	8133384	Bay-0	G->A	PreStop	True	agcaacttctatttgctcagg, tcgtacacggttctgttaagc
AT1G23250	1	8255344	Rrs-7	G->A	PreStop	True	ttgttgttgttgttgatcgc, agaaccttacgatttcatcgg
AT1G23300	1	8265079	Got-7	C->T	PreStop	True	gagacttgagggttcttgagg, ggttatagcagcgactgtgc
AT1G23450	1	8326560	Br-0	C->A	PreStop	False	ctggtttggtggaaaaagc, agtgtcaatcatgtggtctgg
AT1G23560	1	8352793	Cvi-0	A->G	RevStop	True	tgcagagaagttccacacg, gccactttcttggtacatacg
AT1G23590	1	8360504	Got-7	G->A	PreStop	True	tctctccaaagttttccttgc, acactacccccatcactaacc
AT1G23670	1	8373651	Rrs-10	C->T	PreStop	True	gatcttcgattagagccttgg, acagagaagcacgtgaggg
AT1G23670	1	8375769	Sha	G->A	PreStop	True	tcgttggtttgtgatcttacc, tactcctccaccattacctcc
AT1G23770	1	8405903	Cvi-0	G->T	PreStop	True	ctccacaacatgaacacttcc, aaattcgggtatagagggtcc

AT1G24150	1	8549508	Sha	T->A	Met	True	aactactctgctcttacaggcg, cacagccctcaagatatttcc
AT1G24250	1	8589214	Bor-4	A->T	PreStop	True	aagagcttaagcatttcccc, tggaagatgtaaaggcatacg
AT1G24490	1	8679693	C24	C->T	PreStop	True	tctatgtacaccgatcttggc, tgttgtgtcgatagaagtggc
AT1G24490	1	8681651	Lov-5	G->A	PreStop	True	tttgttaagtcttggagctgg, gaactttggtggagaaaacc
AT1G25310	1	8874160	Bay-0	C->T	PreStop	True	caaacaaacgaagaactgagg, aaggaaattacacccaactgc
AT1G25410	1	8914954	Tamm-2	C->A	PreStop	True	accttaggtcagagcagatcg, tctcattcttctctccttccc
AT1G27490	1	9546696	Ts-1	C->T	SD	True	aagaaacgagcaagattgtcc, tcacgttgattgattctaggg
AT1G27570	1	9575585	Van-0	A->T	PreStop	True	acccagtttaaccgaactcc, gtaaaaccaggtacgaaaccg
AT1G28020	1	9769424	Bor-4	G->T	PreStop	True	agacggtttcatttgatctcc, atcaaacttgagtggcatacg
AT1G28500	1	10019611	Bor-4	G->C	SA	True	gctattcctttgcagaaaacc, aagagaaatccaatcagtggc
AT1G29355	1	10275525	Van-0	G->C	RevStop	True	ggaagaagaggaaaatcatgc, atctggaaaaggagaaacacg
AT1G29480	1	10318093	Nfa-8	A->T	PreStop	True	aacgtatttctccttgtgcg, agacactctttcgaagaagcc
AT1G29580	1	10338606	Rrs-7	A->G	SA	True	aattcgagagcaagagaatcc, gttaacacttgtcgaaatggc
AT1G29730	1	10401966	Rrs-10	A->G	SD(non)	True	ataaagcttcacaaggttcgg, agaacgaggtattgaattcgc
AT1G29870	1	10458751	Rrs-7	G->A	PreStop	True	tgaatgaatcctcaccttacg, ttgagcaccaagtttcgc
AT1G30000	1	10510092	Rrs-7	A->G	SD(non)	True	agattagtgagggaaaatcgg, taggtatgcttcttgcctgg
AT1G30020	1	10516322	L*er*-1	T->A	PreStop	True	gagctctttgcttttgactcc, tcacagagttatgatgcagcc
AT1G30160	1	10606449	C24	A->T	SA	True	tcttctgcaataatgtctcgg, catagggttcaaatagtcgc
AT1G30170	1	10608668	Rrs-7	A->T	PreStop	True	gagggtgactccacaaagc, tcttcttggacacacaagacc
AT1G30690	1	10887810	Lov-5	G->A	SD	True	tccttctctgcataaagctcc, cagaatccaatttcaactctgc
AT1G31270	1	11178339	L*er*-1	C->T	PreStop	True	atccatgtggatgaggtatcg, ggttcctctcctctacacgc
AT1G31530	1	11282402	Lov-5	C->T	SD	True	tgagcttccatgtttctatcg, cgagtcgtctcttacaacacc
AT1G31790	1	11395138	Cvi-0	T->A	PreStop	True	caatctgactcaaatccgagc, gaaaggtgctttcaagattcc
AT1G32140	1	11562956	Rrs-7	G->T	PreStop	True	cttcttctctttccttttgcg, tggacagattcttcctctgc
AT1G32140	1	11564656	Rrs-7	A->T	PreStop	True	actacgagcttgcttgagtgg, aagtaaaaccctagtgaggaagg
AT1G32390	1	11688605	Est-1	T->A	PreStop	True	catacaggagtttggttcgc, ttggaggatgctaaggacg
AT1G32480	1	11741855	Ts-1	T->A	PreStop	True	taactccaaatctctatgggc, caatgtagtgccatttattccc
AT1G32850	1	11904086	Est-1	T->A	SD	True	aaagaaaaggtgtcttctcgc, gaccattagataaggtccctcc
AT1G32880	1	11914302	Br-0	A->T	SD	True	ccatccagacactttaagatgg, tcgatgaaagtttccctaagc
AT1G33390	1	12103921	Lov-5	C->T	SA	True	ttgtaatagcgacgatacatcc, gtacgattatggcaagtgtgg

AT1G33530	1	12160729	Ts-1	C->A	PreStop	True	tgggatagatgttgtgacagg, actggtaaggagaaaccaagg
AT1G33540	1	12162329	Ts-1	C->G	RevStop	True	gtatctagcgataaccggtgg, gacatttgccactatcaaagc
AT1G33600	1	12182184	Sha	A->T	PreStop	True	acatggttcttcaacctagcg, tcttgaggtttaaagtgctgc
AT1G35610	1	13143486	Bur-0	C->T	PreStop	True	gccttcgtagaagatcaaacc, cccatttttatctccacatcc
AT1G35610	1	13143979	Got-7	G->A	PreStop	True	ggatgtggagataaaaatggg, catgaaaatgtttgggatgg
AT1G35770	1	13275302	Got-7	G->T	PreStop	False	aaaacgagatattacctccgc, tactcggaataggaaaggagc
AT1G35770	1	13274360	Lov-5	C->G	SA	True	caccctaccaaatccattacc, gtccatgattcctaggtgagc
AT1G35860	1	13333426	Tamm-2	C->A	SA	True	ccgaactgaagagacaaagc, accaggtctagttttccatgc
AT1G36230	1	13613685	Tamm-2	C->A	PreStop	True	taaaacaagtcgcaacctacg, aagggatgttagggttaatgc
AT1G36920	1	13984917	Br-0	G->A	PreStop	True	tctcactactgctgaaggtcc, tacttacgcttcttgaaaccg
AT1G37037	1	14069393	Nfa-8	G->A	PreStop	True	caaagacaaagactgaaacgc, cgagagtgattacaatggagc
AT1G37150	1	14177689	Br-0	C->T	SA	True	catagttttcttcggaccagc, acattcacattgaggtttggg
AT1G42460	1	15916354	C24	G->T	PreStop	True	cttaagatcaacacacaatgcc, ggtcactctcgaagctaaagg
AT1G43760	1	16531997	Br-0	G->A	PreStop	True	cttacatcatcatccatgcg, ctccttcctttcaactcatcc
AT1G43760	1	16531835	Rrs-10	T->A	PreStop	True	gagagcagagagacgagacg, ctgccaagaagtgttgtaagc
AT1G43920	1	16662874	Bur-0	A->T	PreStop	True	ccgttcttccgttgtattagc, caatgcatacaatacaagctcc
AT1G44880	1	16958421	Tsu-1	C->A	SD	True	ctgattgtttgtcaggtttgg, agcatctcattcgttgttagg
AT1G47270	1	17329486	Est-1	A->T	PreStop	True	tcctctctttctcgtttctcg, tgttcagactcaatcctttgg
AT1G47660	1	17537619	Tsu-1	A->G	SD(non)	True	ggtgctgagcagagttatatcc, tcatgtcgcagttccagc
AT1G47800	1	17604506	Br-0	T->A	PreStop	True	aggtgtttacgttatcttccg, ccctcacatcaaatctcacc
AT1G48060	1	17732600	Cvi-0	A->T	PreStop	True	gttcaaacctttacgcaaacc, tcttttcaccctctaaaaccc
AT1G48090	1	17749200	Lov-5	C->A	SD	True	agagattcagaagaagcctgc, caaagctacctccagatttcc
AT1G48730	1	18024007	Rrs-10	A->T	RevStop	True	tttggcccattttatcagc, gagatcgtgagtcaagaggc
AT1G48880	1	18085308	Rrs-10	T->A	PreStop	False	gaaagggttcaattgcttagg, cttacgtttcgaaaagcttcc
AT1G49015	1	18138734	Rrs-10	G->A	PreStop	True	tcaagaaccagcctaaaaagg, ctacatctcaagcttatccacg
AT1G49250	1	18227924	C24	A->C	RevStop	True	agacaagaatccagaggaagc, tgagtggcagatgagataacg
AT1G49640	1	18379703	Bur-0	G->A	PreStop	True	cagtgtccttcctcttcttcc, tacgacgattcatggtctgc
AT1G49920	1	18486977	Got-7	C->A	PreStop	True	tggatagactgtgaaaatgcc, taagagagacggagaaggacg
AT1G50870	1	18858990	Cvi-0	T->G	PreStop	True	agttttagccaacaatggagc, actgtctcatggtagggttcc
AT1G50870	1	18858988	Fei-0	G->A	PreStop	True	agttttagccaacaatggagc, actgtctcatggtagggttcc

AT1G51480	1	19097902	Van-0	C->A	PreStop	True	ggtttccgatatggagttagg, ttggagattaaggtggagagg
AT1G51520	1	19112148	Van-0	A->T	PreStop	True	gtttggtgaaacaatagcagg, gagagcacaacaacaacaacc
AT1G51530	1	19115127	C24	G->T	PreStop	True	acaatccctccaaagtaatgc, gacttgagatgggaaatgagc
AT1G52060	1	19362907	Van-0	G->C	PreStop	True	tatcggagagaaataaacccc, acaaggagggagagacagagg
AT1G52590	1	19593295	Ts-1	G->C	PreStop	False	agcgagaacttacagaggagc, gagaccatgatcgaaacagg
AT1G52615	1	19604329	Tsu-1	T->A	PreStop	True	agattgggtcttatggtttgg, cagtcatattggcgatagtgg
AT1G52770	1	19660604	C24	A->C	SA	True	tatctctcgtcgaagctctcc, tgtgaaccagaaggattaggg
AT1G52810	1	19671271	Van-0	G->T	PreStop	True	atgcttctgaagaggttttcc, tttctgtctgtatgagcctcc
AT1G53265	1	19865082	Got-7	C->T	PreStop	True	gtgggagtgacattaaaaacg, cacgaaatgaaacaatctcg
AT1G53265	1	19865081	Lov-5	C->T	PreStop	True	gtgggagtgacattaaaaacg, cacgaaatgaaacaatctcg
AT1G53930	1	20144081	Br-0	A->G	RevStop	True	tggcgagaagaatagattatcg, ggagaatcgtcttagaggtgg
AT1G53950	1	20151365	Cvi-0	C->T	SD	True	aacctcagatgtggacaaacc, ctgagtccaactccagatacg
AT1G53990	1	20155881	Fei-0	C->A	PreStop	True	caaaagggttttctcaagagc, aaggactgttgattctttcgg
AT1G54100	1	20199902	Lov-5	G->T	SD	True	tagctcttctttgacgacagc, taatccccttagctttgttgc
AT1G54170	1	20225041	Br-0	G->T	PreStop	True	aaagaaatgagatcagcacg, ttaaaacagtacccacaaccg
AT1G54430	1	20320912	Br-0	A->C	PreStop	True	ccataccatcgacatatagcc, agtgcatcacaaatcatctgg
AT1G54430	1	20320953	L*er*-1	G->A	PreStop	True	ccataccatcgacatatagcc, agtgcatcacaaatcatctgg
AT1G54760	1	20437795	Cvi-0	C->T	PreStop	True	ggtcacattatctaagcgtcg, tagagcccttcacacttttcc
AT1G55010	1	20520923	Br-0	C->A	PreStop	True	aagttttgtaccaccattaccc, tcctgattttggatattgcg
AT1G55380	1	20681942	Nfa-8	G->C	PreStop	True	tgagatagtggtaggtggtgg, caaaggcaagcaattaaagc
AT1G55535	1	20737467	Bur-0	A->G	Met	True	ctcctagtgatccgattagcg, taatttggcctcaatcttgg
AT1G55650	1	20802142	Lov-5	C->T	PreStop	True	gagaaaagtcgtgaggatgg, gatgttactttcagaaacggc
AT1G56460	1	21153070	Est-1	T->A	PreStop	False	accttcccagaagaacttgg, ctgggttacttggttaggtcc
AT1G58235	1	21584285	Br-0	G->A	PreStop	True	acggagaagctagacaagagc, tctaaagtcaccagaaatccg
AT1G59620	1	21907322	Nfa-8	C->T	PreStop	True	ctattgggtacacgaaaggc, atttgcttaccaagctcttcc
AT1G59620	1	21908208	Rrs-10	C->T	PreStop	True	ggagatttgaaacatgtcacttgc, gggagagagaggtatttcagc
AT1G59660	1	21929217	Fei-0	C->T	PreStop	True	ctctcagtaggacttggaggg, aaaccctgttgtgtttgtgg
AT1G60380	1	22250997	Sha	G->C	PreStop	True	gtctcctccgactttatcagc, gtttgagatttgcttcatccg
AT1G60540	1	22307503	Cvi-0	T->A	PreStop	True	tgatcattgaaagcatcgg, atactgtgttgcaggatctgg
AT1G60540	1	22306407	Van-0	A->T	PreStop	True	agagttctgatcaacaatggc, ctcagaagacaacctcacagc

AT1G60630	1	22338795	Tsu-1	G->C	PreStop	False	cttcaatacctctctcatcgc, gaaaatagcagaggacttggc
AT1G61500	1	22696545	Br-0	A->C	Met	True	gggaatggtatccttgaacc, gtttctatatgaggccaaccg
AT1G61700	1	22791564	Sha	A->C	SA	True	tgtaatctttagcccgtttgg, cggatttctccatagctgc
AT1G61730	1	22798151	Sha	G->C	PreStop	False	gtttaaagcgattgtcttccc, tattcccaaacacttttgcc
AT1G61870	1	22869757	Bur-0	A->C	PreStop	False	tattcggctttatccctttcc, aagatccagatcgtatcctcg
AT1G63190	1	23435597	Est-1	G->T	SA	True	caaaacaaacaatacgagggc, gtgaagtcgatatcaaaaccc
AT1G63350	1	23500878	Rrs-10	G->T	PreStop	False	cctctttagacaccacaaccc, gaaagctcaaaagagaggagg
AT1G63350	1	23498602	Tsu-1	A->T	RevStop	True	gcaatagtagagttcagtgccg, atctgaaaaagcttccactcg
AT1G64020	1	23755420	L*er*-1	T->A	RevStop	True	cttgtaacggatccatgagg, tcctcggatcttcttctatcg
AT1G64030	1	23756801	Nfa-8	A->T	PreStop	True	attagatcccgaacaaggtcc, gtcctcaggcttccttacc
AT1G64100	1	23796842	Sha	C->A	PreStop	True	atgaaggaatgcaacttctcc, tgcacttttccacagaacc
AT1G64600	1	24002245	Rrs-10	G->A	PreStop	True	gaatcatattccctccattcc, gtcttgtgcactgtttgttgg
AT1G65370	1	24288903	Tsu-1	C->A	PreStop	True	atcagtatcatccatcaaggc, aagggatttatcgtggaagg
AT1G65510	1	24363424	Rrs-10	G->T	PreStop	True	cctctcaagttaaagtcgtcg, tttgacgatctacacaatcgg
AT1G65990	1	24575691	Est-1	G->A	PreStop	True	aatttccacaagcatctagcc, ttgatgtgtctccacactgg
AT1G66020	1	24582402	Ts-1	C->A	PreStop	True	ctcagtgtacctcagcattgg, tatagcctcaggaagagacgc
AT1G66360	1	24755730	Nfa-8	C->T	PreStop	True	ggtcatttgaatttcgtaggc, aacctgtccctatcgtacacc
AT1G66380	1	24762151	Got-7	T->A	RevStop	True	tcgtatgtatacgtaggtggtcc, tctctccatcgaacaaattcc
AT1G66490	1	24813247	Est-1	C->T	PreStop	True	ctctttctttccttttgccc, tactgctgagagatttggacc
AT1G66650	1	24863921	Nfa-8	A->T	PreStop	True	gtcccaatgatatgatctaacg, cgctacacacaaagtgagtcc
AT1G66950	1	24981902	Ts-1	A->G	Met	True	atcctcaattatatctgccgc, tccatgtcatcttcctccc
AT1G67270	1	25188660	Nfa-8	C->A	PreStop	True	atttccagcttcttcagttgg, ggatattcctcagtctatgcg
AT1G67900	1	25471280	Lov-5	A->C	SA	True	atgaacttttcttctggtggc, cttcagaggacacacatctgc
AT1G68585	1	25761087	Fei-0	T->C	RevStop	True	tttgttctttcctctcacgc, aatctctggcaacaaagtgg
AT1G68740	1	25818109	Bor-4	G->A	SD(con)	True	tagtttatcctcgcttttcgc, taatctttgagtgttgggtgg
AT1G70620	1	26631437	Rrs-10	G->A	SA	True	tgtgggtttctttgtaactcg, ctactccacttgtggtatggg
AT1G71150	1	26832176	Cvi-0	A->T	RevStop	True	agcttggagcttgtgtttacc, agctgattcatgcattgtagg
AT1G72060	1	27122397	L*er*-1	T->A	PreStop	True	caggagaagcagaaatcttcc, gaaatcgatgagtgggagg
AT1G72250	1	27198554	Sha	G->A	SA	True	ctctctccttcatatttcgcc, tttcctctcttttctcctgc
AT1G72300	1	27224626	Lov-5	C->T	Met	True	cgaagtaacacgattttcagg, tggtctgactattccctttgg

AT1G72320	1	27236199	L*er*-1	C->A	SA	True	ctttcctgaaaaagatacaccc, tttggggtctgtaaatttgg
AT1G72450	1	27277998	Ts-1	C->A	RevStop	True	acgcaaatccatcactgg, acacggtagtcatcttcctcc
AT1G73570	1	27656670	Got-7	T->G	PreStop	True	tactatttgaggttggggtgg, atctccaataagcaatgcagc
AT1G74170	1	27895323	Nfa-8	G->A	PreStop	True	tatgcattgaaccaacaacc, ggtaatcctcttctctgtggg
AT1G74170	1	27897749	Tsu-1	G->A	PreStop	True	caattctttaccagaaagggg, ttcaatagcaggatctttccc
AT1G74280	1	27933741	Rrs-10	T->C	SD(non)	True	gttgtgattgtcattgttggg, tgacacactgttagaggtccc
AT1G74310	1	27942450	Ts-1	A->T	PreStop	False	gctttatccttctccctttcc, caagcccatgttagctagagg
AT1G74420	1	27971690	Rrs-10	T->A	PreStop	True	aatctcagcttgagccatagg, ccttctgagttgttctgatgc
AT1G75790	1	28458794	Tamm-2	C->T	PreStop	True	ctctgatcattcctttaaaccg, agatatggatttggagcttgg
AT1G76170	1	28590296	Van-0	A->G	SD(non)	True	tgtttcggctatatcatctgc, tattgatgaagggataacggg
AT1G77250	1	29026055	Br-0	A->T	PreStop	False	tgtggtaatacttgtgtgggc, ttgaagcagttctacactgggg
AT1G77300	1	29053696	Rrs-10	A->T	PreStop	True	tctgtaaagcacttcctttgg, atgttgatgattgagccatcc
AT1G77410	1	29095381	Bur-0	G->A	PreStop	True	gaatctcccattaacaaaggc, gaatgtagccttcaacactgc
AT1G77880	1	29291787	Rrs-7	C->A	PreStop	True	aagcaatgactcaaacagtgg, agaagaactgtgttggtcgg
AT1G78640	1	29586879	Nfa-8	C->A	Met	True	tttaggaacgtcgacaatagg, tagggggatagagttttgacc
AT1G78840	1	29645602	Bur-0	T->A	PreStop	True	tcacacatagaaatgcttccc, agcattacatcattgctgagg
AT1G79670	1	29981783	Cvi-0	C->T	PreStop	True	gttgaaatgtgtgtgtaatcgc, tctaaaagggaagaaacgacc
AT1G80310	1	30201202	Bay-0	C->T	PreStop	True	gcttccaaagacatgaactcc, ccagttttggctcttttatgc
AT1G80960	1	30422814	Rrs-10	A->T	SA	True	gagtaggaaaagagcattggg, ttgatcatcatcttctcgacc
AT2G02440	2	639159	Lov-5	G->A	SA	True	ggtgattgtgcatgtcttcc, tcctccactctgtttcatgc
AT2G02710	2	758925	Bur-0	T->A	RevStop	True	ctcttgagtgcacatactcgg, catctcaccagttcgtaatgc
AT2G03540	2	1074480	Tsu-1	C->T	PreStop	True	tcagatcgaactactcaacgg, gagatcgtacatgtgggtgg
AT2G04410	2	1534270	Got-7	A->T	PreStop	True	ggaagaagaagaagaagagatgc, atcttgaacagtggatagggc
AT2G04580	2	1599855	Bor-4	C->A	PreStop	True	aaagtaagtcactcttcgggg, gcagatgaagaaagcataagg
AT2G04580	2	1600097	Bor-4	T->C	SA	True	caagacaaccatgtatgctcc, gatgtagagcagaagaagagacg
AT2G04930	2	1734151	Got-7	C->T	PreStop	True	atcatcaacaatcaccactcc, tttacagctaagcaacgaagc
AT2G05420	2	1984165	Bay-0	T->G	PreStop	True	ttcactctgatcaattcgtgg, tttagttgggagagttggtcc
AT2G05970	2	2308681	Tsu-1	C->T	PreStop	True	ggtagcccatattaacaatactacg, ctattgttgccaaagacatcc
AT2G06500	2	2581973	Tsu-1	C->T	SD	True	cagtataaccgctgtataatccc, cgtgcttcaaagtacttctcc
AT2G07170	2	2977827	Cvi-0	C->G	PreStop	True	ttagagtatgcagtggagggg, gtcaagaagctcaacaagtgc

AT2G07320	2	3039630	Bor-4	C->A	PreStop	True	ctcaaatgtaatagcagttttcccc, taaatagggaaagcgcatgg
AT2G07320	2	3041295	Tsu-1	G->C	PreStop	True	gggacattgtttgtttaagagc, caaacccgatataagtgaccc
AT2G07760	2	3585801	Nfa-8	G->A	PreStop	True	agaagacatgacttggaatcg, tgaatcttacttcttcgctgg
AT2G10440	2	4022430	Van-0	G->A	PreStop	True	ggatgaaacgttagactcttgc, atttcctccttctcaaacagc
AT2G10850	2	4284188	Est-1	G->A	PreStop	True	cgtgtctaccttcatagtggg, aatgtcaagaggtgaaatccg
AT2G10965	2	4331981	Bur-0	C->A	PreStop	True	gagttagggtgaacaagctgc, ctatctcccttcatctttccg
AT2G10980	2	4344364	Bor-4	G->A	PreStop	False	gttcgatgatgttgaacaagg, gtgatgtacatttgaatcacgc
AT2G11360	2	4538583	Fei-0	C->G	Met	True	tgaaaacaacatttggaggg, tggcaaattgtactgttacgg
AT2G12875	2	5296636	Got-7	C->T	SD	True	ttatcatctccggattcttcg, gaagaagggaatgattgttgg
AT2G13430	2	5598247	Est-1	C->T	PreStop	True	agggacaatcatcaatcaacc, tgatgaattctcttgtcgtcc
AT2G13500	2	5635225	Bay-0	G->T	PreStop	True	aacatggtcgatgtgttatgg, agcaatctgtgtagaagtggc
AT2G13510	2	5637217	C24	G->A	PreStop	True	tgagtggttaacgcttcttagg, ttgaactccacctccaagg
AT2G13975	2	5872137	C24	C->T	PreStop	True	gtgcttgtgtttgtttcgg, ggatcatgtgttcaaatggg
AT2G14000	2	5892104	Ts-1	G->A	PreStop	True	tccttgtcaatgctatcatcc, cttattccattgttccttggc
AT2G14020	2	5901217	Van-0	G->C	PreStop	False	aaatagagagtcgcccatagc, gaatccaataagttcgcttcc
AT2G14710	2	6306368	C24	G->A	PreStop	True	acaagacgttcatcaacaacc, ttatgaatgtggtagcacaagg
AT2G15420	2	6731960	Bor-4	G->T	PreStop	True	caagctcacggattgtcg, ggtatttcagtcgattggagc
AT2G16220	2	7038304	Br-0	T->A	PreStop	True	gtataccttcctgtgttgggg, aatctccttcttcttcgttcg
AT2G16575	2	7191822	Rrs-7	C->A	PreStop	True	aggtatggacacaaccactcc, ttatcgtcaaaatctgagggg
AT2G16810	2	7295061	Bor-4	T->A	PreStop	True	tgcccatatattatcagtgacg, agtcagagctttggattttcg
AT2G17060	2	7433909	Rrs-7	G->T	PreStop	True	atttcaagtcacagatggtgc, aaaataaaagccactcgtgc
AT2G17670	2	7682548	Bur-0	T->C	SD(non)	True	ttttgcacactgagtaaaggg, aagcatctgacaaactcttgc
AT2G17860	2	7769387	Bur-0	G->T	PreStop	True	cgtcagagacgctaactgc, gctgaaagtaacggttgaagc
AT2G18190	2	7922578	Tsu-1	C->T	PreStop	True	aagattgatacatccccaacc, tcatcggatttaagaggaacc
AT2G18920	2	8205161	Got-7	G->A	PreStop	True	aattcttgctcatcgaaacg, gcagaaaaatgtggtttgacc
AT2G19150	2	8314499	Cvi-0	G->A	PreStop	True	gcttggaaagcaaaaggc, ctcctaagagttgaaacttgtgc
AT2G19600	2	8489425	Bur-0	C->T	SD(con)	True	tcttagctattggctgtttgc, gctagcagaatgtcaacatgg
AT2G19910	2	8603611	Sha	T->A	PreStop	True	ccggttcatcttcaaaatacc, ttttcacggtttacagagacg
AT2G19920	2	8609304	Tamm-2	C->G	RevStop	True	taggttacttgcctaacacgc, ccagaaatacaagcaggtaagc
AT2G19980	2	8634873	Nfa-8	A->T	PreStop	True	tctttcaactcaaaacacgc, gtacggagagaatatagccgc

AT2G19980	2	8635135	Lov-5	T->C	SA	True	aataaggcttctccgtaaacc, agctcttgttttgtttgaggc
AT2G20250	2	8742833	L*er*-1	C->A	SD	True	caccacttggaactgtgtacc, accagacaaaggtttttgagc
AT2G21790	2	9303732	L*er*-1	C->G	PreStop	False	agtcccgtgagtaaggtatgc, cccatatcagttaggtcatgg
AT2G21800	2	9307900	Got-7	G->A	PreStop	True	cctctagggtccaagattcc, cttgttaaccgaacaacatgg
AT2G22350	2	9503613	Rrs-7	G->A	PreStop	True	tttgtaatggcagtatggtgg, ctacttctaattcaacccgcc
AT2G22440	2	9536996	Rrs-7	A->G	RevStop	True	aaaagaggatgattccactcg, ccatttaggaaccaatgtgc
AT2G24600	2	10460909	Got-7	G->T	PreStop	True	ctgctagttcccaaattctcc, caatttcgccttatcaagtgg
AT2G24600	2	10459552	Sha	G->T	PreStop	True	cttcaagtaaatcttgtgccg, ggtttctgtagggtttatggc
AT2G24630	2	10479412	Got-7	G->A	PreStop	True	cagatgacccaaataggaagc, acaaccgtcgaggatatgg
AT2G24650	2	10490755	Bur-0	C->T	PreStop	True	gccaactaaaaacttacacgc, tcgtgacattagcacttacacc
AT2G24830	2	10585031	Cvi-0	G->A	PreStop	True	tgagcgtactctgataatgcc, tcttgtcgttcatcacatgg
AT2G25360	2	10811330	Rrs-7	G->A	Met	True	agggttgttctacagtcgtcc, cttcttaatcttccctcgacc
AT2G25450	2	10838380	Sha	G->A	PreStop	True	aacctcaggtaagtcctgtgc, ccagtgagttaaaagcattcg
AT2G25590	2	10898977	Got-7	A->T	PreStop	True	tcacggttcacaaagtttacc, ttcacgtacaattcaacaccc
AT2G25710	2	10960456	Bay-0	C->T	SD(con)	True	tataactcatcggtttggacg, tttcactgcctcagttacagc
AT2G27050	2	11555069	Nfa-8	C->T	PreStop	True	ccatgtacgacagaaatgtcc, tatagatgagtgtttggtgcc
AT2G27120	2	11595997	Ts-1	G->A	PreStop	True	ttattggctcgagaaaaagg, gcatcaatctgattacaaggc
AT2G27760	2	11833886	Cvi-0	T->C	SD(non)	True	tgacatctacaaaccaggagc, aaagtttgttctccaagtggc
AT2G28520	2	12222608	Cvi-0	G->C	RevStop	True	agttcttcttcttgcctggg, cttctgacatttactactacggc
AT2G29525	2	12647322	Bur-0	T->A	RevStop	False	acacagttgacattgttgttgc, tgtgcagttagaatggtctgg
AT2G29710	2	12707196	Br-0	A->G	RevStop	True	caatgtatgcagagcaacagc, ccataacagaagaaatgcagc
AT2G29720	2	12707671	Tamm-2	A->G	RevStop	True	gctgcatttcttctgttatgg, gaccgaatcagttgaaggg
AT2G29780	2	12725796	Bay-0	C->A	PreStop	True	gatcataaaccaaaccacacg, attccctcctcatagtttccc
AT2G30430	2	12975488	C24	A->G	RevStop	True	acacatttgacagcatccg, gtgattgtggacaagaaaagc
AT2G32050	2	13644957	Tsu-1	C->A	RevStop	True	ctgtgtattcagttccaaccc, aagaagaagttgcaaaggagg
AT2G32340	2	13740399	Rrs-7	A->T	PreStop	True	tgttttggtttgacaggaacc, atgccatagaggcatcttacc
AT2G32490	2	13799246	Rrs-10	T->G	Met	True	ctccatcatcattcttcatcg, gtagcgaaggctgtaaattgc
AT2G32910	2	13966727	L*er*-1	G->C	SA	True	atttggaaacctttcggg, tctacaacctcattcccatcc
AT2G33160	2	14063204	Nfa-8	A->T	PreStop	True	ttggagtaaaggatatcgacg, gcttatagagtccggtgaacg
AT2G34240	2	14466354	Tamm-2	T->A	PreStop	True	tcagtcaaagccaaagaaacc, aactaacactcccccattttgc

AT2G34850	2	14712894	Bur-0	T->C	SA	True	tcttctgccatctttttagcc, gatgcaaatgctgtatgatcc
AT2G35140	2	14823886	Cvi-0	A->T	RevStop	True	gacgtaatgaatgtgactcgc, gtttctgaaccaaacacatgc
AT2G35330	2	14877593	Cvi-0	G->T	PreStop	False	aatgtttcagttcagtgctgg, cacgccagtagtttctttaagc
AT2G36340	2	15244083	Van-0	G->T	PreStop	True	aataaggatgttccttgtgtgg, atcccagaatcaagtgtgtcc
AT2G36650	2	15367328	Rrs-10	G->T	PreStop	True	gagatggaggagctatgaagc, ctcaactcttggatttctccc
AT2G37680	2	15811092	L*er*-1	G->A	PreStop	False	tgttagtccacaatctgtgcc, gtcataacattctgggaaggg
AT2G38150	2	15991310	Tamm-2	C->A	PreStop	False	tggacacaatcttacacaaacc, gatcttcgagaaagatcaccc
AT2G38160	2	15994431	Nfa-8	C->T	SA	True	ttgatttctgagcagagttgg, gcctaaagctaagtccactgc
AT2G38590	2	16150498	Tamm-2	T->A	PreStop	True	acgttgatccatctaaagcg, ctttcctttctttagcacacc
AT2G39650	2	16535174	Cvi-0	A->T	RevStop	True	ccttggctgttattgtaaccc, cagcgaaattctccttaaagc
AT2G41430	2	17277166	Br-0	C->G	PreStop	True	ggatggtttctatgacaacgg, tggctaaagatacacagaccg
AT2G42240	2	17603969	Est-1	G->T	Met	True	atgctattatcctacgcctgg, aagattctcagtctcttcgcc
AT2G42245	2	17605915	Rrs-10	C->T	PreStop	True	atttccaggtacagctcttgc, accccaacaacactattctcc
AT2G42270	2	17615083	Ts-1	G->T	PreStop	True	agttggagaaaaacgatctgg, agtagctctgtaggtggtggg
AT2G42340	2	17642660	Lov-5	G->A	PreStop	True	tttgattgctcaagaatcgg, aacaaaaccggaaagtctagg
AT2G42370	2	17650547	Ts-1	A->T	PreStop	True	tctgtttttgatgacggagc, tgctgcttacgtttcttatcg
AT2G42590	2	17739199	Cvi-0	C->G	SA	True	taagagtgtcctgagacaggc, tgaaatagcatctggaagacg
AT2G42630	2	17765784	Sha	A->G	SD(non)	True	actcacagaatcagcaaatcg, aagcgacatctttagctttgg
AT2G42960	2	17875674	Sha	C->G	RevStop	True	cagtcatcatcactccacagg, tgcttgaatctgatgaacacc
AT2G43270	2	17995864	Nfa-8	T->C	SA	True	accgtgaaccaactagactcc, aactatacctcacttcctttccg
AT2G43730	2	18132087	Van-0	G->T	PreStop	True	aatgtgtcattcctccatcg, acactaatcaggggaaacacg
AT2G44280	2	18310747	Rrs-10	G->C	PreStop	True	aaccgccaaaaacagagg, agaagatccatcgacaaaacc
AT2G45135	2	18615804	Sha	A->T	PreStop	True	tggtactgttagacacctcgg, tttttggagtagacatcaccg
AT2G45920	2	18906854	Nfa-8	C->A	PreStop	True	gcaaggagttctgtatcgtcc, catatacaacaagtagcaggcg
AT2G46480	2	19084627	Bor-4	A->C	PreStop	True	ctgccaaataaacccagtagg, aaatactggctagagcacacg
AT3G01260	3	81306	Bor-4	T->C	SA	True	catcgctagactctttcgtcc, ggtgcttctatcatgtctctcc
AT3G01620	3	235376	Van-0	G->T	PreStop	True	attttccaagggtagatacgg, cttatagccatactgaccggg
AT3G02980	3	670905	Bor-4	T->A	RevStop	True	gccaatgaagcaaaagagc, atcaacggttctcgaactcc
AT3G03930	3	1011445	Cvi-0	T->G	RevStop	True	gtacatggcatcaagttgtgg, acttcttccttccctagctcc
AT3G05110	3	1426978	Tsu-1	T->A	SA	True	accatgtcacttgaagactcg, aagaagcattagccagagagg

AT3G05450	3	1575033	Br-0	T->A	PreStop	True	ataccttggtttcgatgaccg, aacgcaaataagtgtcacgg
AT3G06010	3	1805810	Bay-0	G->A	PreStop	True	ttgtctcccatgccaagc, ggccccttttgaactatgg
AT3G06110	3	1843941	Est-1	A->G	SA	False	ttggatgagtttattctcaggg, tcctatggttactcagttggc
AT3G06620	3	2064609	Sha	T->A	SA	True	gactgagtcaaataggagggg, ctaacgtcctgctgtttatgg
AT3G07040	3	2227823	Nfa-8	A->T	PreStop	True	aactatcagccacttcttctgc, gactactcgggacatgaacg
AT3G07500	3	2392954	Got-7	T->G	PreStop	False	tgaggacctgcttatcttcg, ctttattacacacgagtctgcg
AT3G07540	3	2406039	Bur-0	T->A	PreStop	True	cagcaacattttgactcttcc, ttcttcaacttctgcatcacc
AT3G07770	3	2483974	L*er*-1	T->G	RevStop	True	aaaactacagcccgataatcc, aatctttccaaaaacacccc
AT3G07920	3	2526235	Fei-0	T->A	Met	True	ggatagctgatgtaaaagggc, aaaattagggttgggaaacg
AT3G08990	3	2744407	Bur-0	G->A	SD	True	gtttgtgcttgtgttgtttcc, tcacattggttacacaacatcc
AT3G10510	3	3275100	Rrs-10	G->A	PreStop	True	tcgttgataaacttgtgagcc, gagaaccaaaagaaacatgcc
AT3G10790	3	3377979	Got-7	G->T	PreStop	True	gacttttctccacctgttcg, tcgttacagctatcttttcgc
AT3G10820	3	3388431	Br-0	G->A	PreStop	True	ccacggtttgatttagatgc, gacctgagaaagttcaaaccc
AT3G10900	3	3410258	Fei-0	T->A	RevStop	True	agagttgagatcaagagacatcc, actctggctaatagagacggc
AT3G11160	3	3496586	Br-0	C->G	SD	True	ctgctgactcccatagactcc, attgggcttaggtatgaatcg
AT3G11380	3	3564995	L*er*-1	G->A	PreStop	True	gagattacaaggcttgatggg, gtgtagccaatcctctgatcc
AT3G11964	3	3801073	L*er*-1	T->A	SA	True	gcttatcaagctcatatccagg, gattcagtgtttcatgttggg
AT3G12420	3	3948142	Est-1	A->G	SD(non)	True	tggatcttgttttcacagacg, gcggatactctccagttatgc
AT3G12430	3	3949589	Rrs-7	A->C	RevStop	True	attattaacagctccgcttgg, aggaagttcgtttcagattcg
AT3G12840	3	4085862	Br-0	G->A	PreStop	True	actcaaaagcttccagactcc, gttcatctattccaagcaaagg
AT3G12850	3	4089920	Br-0	C->A	PreStop	True	agtgtagccatgtaagcatcg, gaaactacgaaggacgaaacc
AT3G13210	3	4245313	Ts-1	A->G	SA	True	agctttccgactacagactcc, cgaatataagcagaaaccttcg
AT3G13370	3	4341673	Est-1	G->A	Met	True	acatccccatttcctagtcc, ggtccacttttataacctccg
AT3G13662	3	4467415	Lov-5	G->T	PreStop	True	ggagaaactcactcatctccg, atatgtagattggcaacaccg
AT3G14490	3	4864456	Van-0	A->T	PreStop	True	gatagccaagtatgctttccc, tgtagaggtctactttggggc
AT3G14650	3	4923885	Got-7	C->T	PreStop	True	agtgtttggcgataaagaacc, agctcaaaggagaatctctgg
AT3G15605	3	5289046	Bur-0	G->A	SA	True	gtttcctgtgttgttggtcc, gatcaacctttctttaacgagc
AT3G15605	3	5288941	C24	G->T	SD	True	atgatgtggtaagtttgctgg, ctggagaagccctagtaatgg
AT3G15930	3	5389613	Lov-5	A->C	SA	True	catacttgcctgatacttccg, tgaagagacatctcaggctcc
AT3G17150	3	5849289	Tsu-1	G->A	SA	True	gcaaaacctcaaatctacaagg, tggtcctcacaataacctgg

AT3G17190	3	5867826	Rrs-7	G->A	SA	True	catggataataacagcatgagc, tgacaggacatcattctctgc
AT3G17265	3	5900458	Bay-0	A->C	PreStop	True	tttgaccagtacaagggttcc, ttctatttgcagtcctgttgc
AT3G17270	3	5902610	Van-0	T->A	PreStop	True	tgggtgtttgacctatattcc, ctatttgcaaacgatggtacg
AT3G17280	3	5903461	Ts-1	A->G	RevStop	True	gcttgattggtattgttgtgc, gtgtaacgagtttcctgttgg
AT3G17400	3	5955367	Got-7	C->T	PreStop	True	acgttgacgcaactataaacg, atccttctttgtgcatttgg
AT3G17450	3	5973039	Est-1	C->A	SD	True	aggatcaaggttgtcttcacc, gttaacagtcaatgccagagc
AT3G17620	3	6027300	Est-1	C->T	PreStop	True	aaaactctacttcccagtcgg, atctaacgagcatcaacctcc
AT3G17670	3	6041269	Bor-4	T->C	SD(non)	True	aactccgagtcactcactgc, ccaattttgctactgatctgc
AT3G18485	3	6341925	Ts-1	T->G	RevStop	True	cggttgaatatttgtacagagg, cctatgagttcggttaccagc
AT3G18680	3	6429316	Fei-0	C->T	SD(con)	True	ggtttccgtttaagttttctcc, aacaatggagttccaagttcc
AT3G18910	3	6522571	Van-0	T->A	PreStop	False	tttgactctgattcatggagg, caatttctcacagaacatgacg
AT3G18980	3	6546874	Rrs-7	C->T	PreStop	True	atgttgacaagcgagtacagg, ggagtgagatcaccatcagg
AT3G19040	3	6567427	Fei-0	A->T	PreStop	True	acaagcgatcacatttcacc, gtaggtttttgttgtcttgcc
AT3G19070	3	6593779	Bor-4	A->C	Met	True	aaagtaatcagttctctgtacgcc, aaattaacctgcttctctcgc
AT3G19210	3	6653726	Fei-0	G->A	PreStop	True	atgtcttcttggtcaatctgg, catgaacagacggataacagc
AT3G19470	3	6750371	Got-7	G->A	PreStop	True	aagactaggctcgttgtttgg, aactgtaaataatcccacggc
AT3G20080	3	7010326	Lov-5	G->T	PreStop	True	ttctgattcttgggaagatcc, ttacggtttcacacattagcc
AT3G20270	3	7068961	Bor-4	A->T	SA	True	cttcttccagtttccacagc, atagacccattcgaactttcc
AT3G20280	3	7072793	Bay-0	A->C	SA	True	tgcaaatcttactggtcatgg, gagtgtccatttcattgatgg
AT3G20690	3	7232284	Bor-4	C->T	PreStop	True	agagggaaactagaaaccatcc, gttctatccaaacatcggagc
AT3G20710	3	7238750	Sha	G->A	PreStop	False	aggttattttggttgacaccg, gagattctctctagggttccg
AT3G21130	3	7408375	Bor-4	A->T	PreStop	True	gttttggtgcgtgagtatagg, ctgcatagtaataagccgtcg
AT3G21175	3	7424540	L*er*-1	A->T	SA	True	ttagttcacattgatcacctcg, gagattaacccagagaaaccc
AT3G21940	3	7730878	Tsu-1	T->A	PreStop	True	gaagatgtccgaaaaacaagg, cgtactcaaaactctacccccc
AT3G21980	3	7745474	Bor-4	G->A	SD	True	caaatgtaacaacaccgaagg, ttggacatctcctacgaagc
AT3G22421	3	7949368	L*er*-1	C->A	PreStop	True	agactacatatgtctcttcatgtgc, ttagatatacgtccggaagcc
AT3G22560	3	7999383	Rrs-10	T->A	PreStop	True	ccaaaactctttagcgacagc, ttaaccgataagaataggccc
AT3G23080	3	8208932	Fei-0	G->C	PreStop	False	gacttatccatcatagattgcc, ctgttacagagacattcgtcg
AT3G23350	3	8355173	Bay-0	A->T	PreStop	True	gaaaacctccaattcaacacc, gtgatgaagagtcaagaagcg
AT3G23350	3	8354760	Sha	G->T	SA	True	attggagagaagcgtacaagg, cgattaacgaatatgactcgg

AT3G23570	3	8459517	Got-7	A->T	PreStop	True	gggattctctgttacagcacc, ggaagagtagatacgcacacg
AT3G23790	3	8576415	Tamm-2	C->T	PreStop	True	aagaaggtagctagagggtgc, tcatccaacaacagttaaggc
AT3G23860	3	8617267	Got-7	G->T	PreStop	True	tttgagccagagttcatatcg, ggacaacactttcaaacaagc
AT3G23960	3	8658706	Sha	G->T	PreStop	True	ttacctagctatggtcaacgg, ttaaacgcttctaatcctcgg
AT3G24360	3	8840718	Rrs-10	G->A	PreStop	True	tttctcagacacaagacaacg, agagaagccattttgtcagc
AT3G24503	3	8920363	L*er*-1	A->T	PreStop	False	catcagtgacatgactggtcc, ttggtgatccttttgattcc
AT3G24610	3	8978109	Got-7	T->C	Met	True	tcaaacacagaaaccaacacc, gataaccgaaagaatacaaccg
AT3G24700	3	9023037	Rrs-10	G->C	RevStop	True	aattcttgaccacgtctctcc, tagcaaagtaaattcgggtcc
AT3G25420	3	9220101	Bor-4	T->C	SD(non)	True	cgtctgatactcacacattcc, ctttgaatccgaaacatagcc
AT3G25970	3	9503292	Est-1	A->T	Met	True	ctccttaaaagcctcaaaagc, gatgaatgaaagcgttgttagc
AT3G26120	3	9550320	Nfa-8	C->G	PreStop	True	gaaatttccatgcgaggc, tttagaactttgcaggagtcg
AT3G26855	3	9900020	Bor-4	C->T	PreStop	True	cttaggctcttagcgagatgg, atgaaatccatcattgacgc
AT3G26855	3	9900166	Nfa-8	C->T	PreStop	True	tgcaagaagaagagtgtttgg, tcacatcttattcaactgccc
AT3G26920	3	9923611	Br-0	G->A	SD	True	gaatgaaccgaagaatgttcc, gatacaaaccggatgaaaacc
AT3G27260	3	10071386	Cvi-0	A->G	SA	False	aaatggctgatgagtatgtcg, gtattttcatctctttgcgcc
AT3G27540	3	10208077	Cvi-0	C->G	PreStop	True	ggcattgtagcttctgtttcc, gggttatatacgcttcttggc
AT3G27600	3	10225763	Cvi-0	C->A	PreStop	True	ttcagtaatttcaggtggtgg, gttctcggattcaacaagagg
AT3G27600	3	10225295	L*er*-1	C->T	SA	True	caaagtgaggatttctcctgc, tatcttgaatgtctcttgcgg
AT3G27640	3	10233592	Rrs-10	T->C	SD(non)	True	gatctccaaaccaaatggg, tcttccgtactgaaaccaagg
AT3G27730	3	10279205	Est-1	C->A	PreStop	True	agtgacctttttccttgttgc, aacaactggaacgattaaggg
AT3G27800	3	10303245	Fei-0	C->A	PreStop	True	atgatccaacttctttgtgcc, aatgtgccaatctacacaagc
AT3G28040	3	10438333	Tsu-1	G->T	PreStop	False	gtggagaaataccgaaagagc, cttcctgatttcgaagaccc
AT3G28140	3	10471925	Van-0	T->G	PreStop	True	ctacccttttcaattccatcc, ctcttcaccaatctcaaaccc
AT3G28260	3	10535777	Br-0	A->G	Met	True	ttaaccagtcccatacttggc, attctacaaatccccgttcc
AT3G28360	3	10616236	L*er*-1	T->G	Met	True	atttttgctgcttgtctctcc, gggattcttaattaaaacgatcggc
AT3G28370	3	10623573	Bay-0	C->T	PreStop	True	cagacaaaacgaattcagtgc, cggaaaatatctgaacatgg
AT3G28958	3	10984049	Bor-4	T->A	PreStop	True	gattgttgctcgtacaactcg, tcttgagaaacactccatccc
AT3G28958	3	10984148	Est-1	A->G	SA	True	agagaggagacaaaaagaggc, atatgaattcaaggtctcggc
AT3G29050	3	11041938	Tamm-2	T->C	Met	True	tatagacgaaaccgccacg, gtcatcaccttacaatcgtgc
AT3G29150	3	11109664	Est-1	G->C	PreStop	True	caaattaacttggttgtgggg, tacctcagggttttcgacc

AT3G29380	3	11283986	Bor-4	G->T	PreStop	True	ttctgcaaaacacaacagtcc, tagagggaagaagctaaacgc
AT3G29380	3	11284550	Ts-1	G->T	PreStop	True	tttggagacaatctcacaagc, ttaatggaagaagagacctgc
AT3G29750	3	11581711	Bur-0	C->A	PreStop	True	cttataaaaacccataaggccc, ggaagaagtttcatggttgg
AT3G29750	3	11581653	Est-1	G->T	PreStop	True	tgcagaagaatttcagagaagg, atcaaaggaggtaaggactgc
AT3G29750	3	11582678	Ts-1	G->A	PreStop	True	cccaataaaactgaagcttgg, gagatgttcttacaagggttgc
AT3G29790	3	11696101	Rrs-10	T->C	Met	True	atctcactgtcttcaacgacg, tgattcttccttcattcctcc
AT3G29800	3	11722597	Br-0	C->A	PreStop	True	ctcacgtgtcattagaaaccc, gcttttcgtcaaatctcagc
AT3G30200	3	11830412	Ts-1	C->T	PreStop	True	tagctagcaccaccatccc, gccgagtacattgtgttgg
AT3G30240	3	11886421	Br-0	C->T	PreStop	True	caggatatgaaaaatcgcagg, cgaaagtagcgatagtgtttcc
AT3G30240	3	11886674	Rrs-10	T->G	SD	True	ccgagagactctgcttccc, aagtagagagccttgtggagg
AT3G30640	3	12199093	Br-0	G->A	PreStop	True	caatcgatgtgagacgtaagc, gagttagggtattgtcctgcc
AT3G30640	3	12198735	Rrs-7	G->A	PreStop	True	atctctgttgccatcctaacc, gtatcaacagtggatgaacgg
AT3G30770	3	12449944	Bay-0	C->T	PreStop	True	accataggaacttgttttggg, caccgcttattttctgtttcc
AT3G30770	3	12450529	L*er*-1	C->T	PreStop	True	gaaactatggcggattagagg, gttgcaggtacgaacataacc
AT3G32100	3	13100684	Fei-0	A->G	Met	True	attgtttgagttcgagaacgg, tggaagatggagtagtcatgc
AT3G32130	3	13131486	Ts-1	A->T	PreStop	True	ctacccatttggatgttatgg, gaagttggtttgatcttccg
AT3G32150	3	13148284	Ts-1	G->A	SA	True	acgacgaagacttttctgagg, cttggattcagtttgagttgg
AT3G33393	3	14048786	Bur-0	A->G	RevStop	True	tttgaatgaagaggatagccc, ttgcttaggcgaacaacg
AT3G33572	3	14066876	Bur-0	G->T	PreStop	True	tacttctttccaagtgcatcc, tggctacagcaaataatgcc
AT3G42060	3	14264371	Tsu-1	G->A	PreStop	True	cgaagaagaaacacttcctcc, gcgagacttgatttccttacg
AT3G42060	3	14263432	Tsu-1	G->A	PreStop	True	ctcataaaaaccgaccatacg, ggatagaggaacagaggttgc
AT3G42190	3	14379455	Van-0	G->C	SD	True	gtgaccaggtttggttttacc, tatcttcaccatccagagtgc
AT3G42520	3	14645467	Rrs-10	T->C	RevStop	True	agcaggtccggtatttaagg, ggtccacatcttaaatggtaatgtagc
AT3G42580	3	14704541	Got-7	G->T	PreStop	True	tatttcgaaggaggtaggagg, cattttggtaactcagcagc
AT3G42580	3	14702701	Tsu-1	A->G	SA	False	agctgttatgcagcagaagg, tcttgattagacgcagtttgg
AT3G42690	3	14779725	Bur-0	C->G	PreStop	True	tgtgacacatgctgagtttacc, tcttttgcagcttcagttgg
AT3G42690	3	14778365	Fei-0	T->C	SD(non)	True	gaattgtggactgtgtatggg, gctcttgagttcagcaatagg
AT3G42723	3	14851236	Rrs-10	G->A	PreStop	True	tatccattgaccaatgtctcc, tagaaggaaatcatgggaacg
AT3G42786	3	14887585	Bay-0	G->A	PreStop	True	taagagtgttcaccatgctcg, gtttccaaatgggatatgtagg
AT3G42786	3	14888162	C24	C->G	PreStop	True	atactagggagaaggtcgtgg, cgaacaacaaacattcagagg

AT3G42786	3	14887752	Tsu-1	G->T	PreStop	True	agtccgaggaggtattcatgg, agctcctaagatgagtaggcg
AT3G42820	3	14937272	Bay-0	G->A	PreStop	True	aggatattaacctcctactcgg, gatgtgcttgcttagctatcg
AT3G42820	3	14937806	Bor-4	A->T	PreStop	True	caacccttctcttcaacatcg, tctacctcaacgtctgattgc
AT3G42820	3	14932704	Br-0	C->A	PreStop	True	tattagacagagcaacacccc, tccccttcatgcatatttagc
AT3G42820	3	14932204	Cvi-0	C->A	PreStop	True	gtcataaaggagagggcacg, attcactcaggtaatggatgc
AT3G42820	3	14932695	Ts-1	G->A	PreStop	True	tattagacagagcaacacccc, tccccttcatgcatatttagc
AT3G42820	3	14932237	Tsu-1	G->A	PreStop	True	ttcacttcactctgaagaccg, gttttattagcatatgcgggg
AT3G42820	3	14935607	Tsu-1	C->T	SA	True	cacatgcaatgcaactataacc, atcgcgtgtttctttattgg
AT3G42820	3	14933389	Van-0	T->C	SA	True	gaaacaatagaagacgctccc, tcatgtttaccgtatatgcacc
AT3G42820	3	14935539	Rrs-10	A->C	SD	True	cggcaacaacaatgtcagg, atcgcgtgtttctttattgg
AT3G42870	3	14960445	Rrs-7	T->G	Met	True	caaattttggggtcttttcg, gtaccatgcgttttacattgg
AT3G42870	3	14960509	L*er*-1	G->A	SD	True	caaattttggggtcttttcg, gtaccatgcgttttacattgg
AT3G42910	3	14985273	Tamm-2	G->A	PreStop	True	ggaagggataagtttatatcgc, tgtgctaagataatggttctcg
AT3G42910	3	14987696	Tamm-2	C->T	SD	True	ctcctaactcacataacccgc, ccccgtaataatcctctaccc
AT3G42920	3	14997936	Sha	G->A	PreStop	True	ccaaccatctctataatctggc, gggttgtttatgctggtacg
AT3G42920	3	14998143	Bor-4	C->A	SD	True	ggaattgaactgaactgaacc, ggtatcattgattcgttagagg
AT3G43140	3	15123831	Bay-0	T->C	RevStop	True	gaatgggagagactcaaaacc, agtcaagaatcatctcaaccg
AT3G43260	3	15232468	Fei-0	C->T	PreStop	False	agcataacattgaggaggagg, agcctaagaagaagcaaaagc
AT3G43420	3	15356501	Cvi-0	G->T	PreStop	True	gcagatgagaagtaagatgcg, actttggactaacagtatcggc
AT3G43470	3	15403558	Van-0	G->T	PreStop	True	tctttcgtgaacactttctcc, atgtgggagaacctaaagtgg
AT3G43470	3	15404056	Rrs-7	A->G	SD(non)	True	tgtactcatcttgatccgacc, ctcatatggatatagtcaagactcg
AT3G43500	3	15411441	Got-7	A->C	PreStop	True	gtttggctttaatagggagc, cagaccacctgatagacttgc
AT3G43630	3	15550405	Br-0	G->T	RevStop	False	gaggattggtgcaatagtgg, aggtaggtttaacttcttgcg
AT3G43760	3	15659961	Rrs-10	G->T	PreStop	True	agagtcaaaccagaaagaggc, tagttcaaccggtctgttagc
AT3G44040	3	15824807	Bur-0	A->T	PreStop	True	tgttatcttctccaagcaaagc, cacgtctctctcttctctctcc
AT3G44070	3	15839880	Van-0	G->A	PreStop	True	aaacactatcaaaactcccgc, aaactttgtctgataacctggc
AT3G44250	3	15959920	Rrs-7	C->A	PreStop	True	cgatagggctatcaagaaacc, caagattatacttgggaagc
AT3G44350	3	16034069	Bur-0	C->T	SA	True	tgtgtaatagtagtagcattcaccg, tgtggtaaaagagtgtgcg
AT3G44970	3	16444252	C24	G->A	PreStop	True	aagattggatgttaaggacgc, tctctgactcttctattgccg
AT3G44980	3	16452146	Nfa-8	C->G	PreStop	True	cattctacttccacatcagatcc, ggtaaagactctaggcgaagc

AT3G45830	3	16856150	Est-1	G->T	PreStop	True	agatcaaggtgcaacagacc, ttggaataacctctgtttggg
AT3G45840	3	16859170	Cvi-0	G->C	PreStop	True	tgtgaaaaagagttctctcacg, agctgtcttataacatgcttgg
AT3G46610	3	17171299	Tamm-2	A->C	PreStop	False	tttcttcttccctgtaatggc, cttcagttttgacaggacagc
AT3G46650	3	17197401	Bor-4	G->T	PreStop	True	atgtggtcgatggtgtacg, gaaggagttccaatgatttgc
AT3G47110	3	17359882	Lov-5	G->T	PreStop	True	tctagtgagagacaattcgcc, gcatagggaatcttgtaagcc
AT3G47120	3	17362321	Cvi-0	A->C	RevStop	True	gtaaccacaggtcaacacagc, gaacaaaagagatcaggacgg
AT3G48900	3	18144001	Br-0	C->G	PreStop	True	aaatgcaggattgaggagg, cattgcattcatgcttctacc
AT3G49340	3	18305332	Nfa-8	G->C	PreStop	True	taccacattgttgttggtgc, aacttcaggagtcacatctcg
AT3G49340	3	18305286	Tamm-2	T->A	PreStop	True	taccacattgttgttggtgc, aacttcaggagtcacatctcg
AT3G50010	3	18548989	Bay-0	C->A	PreStop	True	tgacttggatgatctagtggc, gaggaagaagtagatccaccg
AT3G50260	3	18646098	Got-7	G->T	PreStop	True	gctcggctcttactctactcc, acagtatgtttgacttgtggg
AT3G51240	3	19036928	Rrs-7	G->T	PreStop	False	tcatcgtctctagtcacctcc, cttgtagcagcaaggtaatgg
AT3G51570	3	19139660	Cvi-0	T->A	PreStop	True	tcttgtcgaccttaagcttcc, ttcaagtcatcgagacaatcc
AT3G51690	3	19187613	L*er*-1	G->A	PreStop	True	cttcaagacaacattttcggc, acaagtgttctctcatggacg
AT3G52690	3	19542390	Got-7	T->A	SD	True	ctttccttggacttatcaccc, tttcccaatggtaacttgtcc
AT3G52780	3	19572863	Got-7	C->G	SA	True	gacacatcatcgtcgttcc, caaggagagaaagagagtgtgg
AT3G53610	3	19889245	Got-7	C->T	SA	True	aatcactgttgtctcaagatcc, tctggttactgtcttgcttgc
AT3G53880	3	19964286	Rrs-7	G->A	PreStop	True	ctgttgttataaaactccgacg, gatatccaatctgcagaaaacg
AT3G53990	3	20001537	Rrs-7	T->C	RevStop	True	tgccttaaattaggtaacgagc, cggagattatggagaaatacg
AT3G54830	3	20322879	Ts-1	T->G	RevStop	True	gagtgatcttctttgcttttgc, acacagacgctgatactaccg
AT3G55660	3	20660308	L*er*-1	T->A	PreStop	True	agctctttctcctcctcttcc, gcaagatccaatacaacaagg
AT3G55670	3	20670463	Rrs-7	G->T	PreStop	True	cacctcttaaaatgagggtcg, catcaaaactttgaaggtgc
AT3G55780	3	20717991	Fei-0	T->G	PreStop	True	ctctgcttcttctttctttgg, agcaaagatcacaagacatgg
AT3G55890	3	20752399	Lov-5	G->A	SA	True	tgagtgagatcttcatgtgtcg, ctcttttcatgtgcaaactcc
AT3G55910	3	20753961	Ts-1	T->A	PreStop	True	tttttctcttcttctctccgc, ataatcatcgtgaagaagccc
AT3G56300	3	20893023	Est-1	T->A	PreStop	True	tggaaagactcaacagtctgc, agatttaggcttggtcaatgg
AT3G56660	3	20998989	Est-1	G->T	PreStop	True	aatctctgttttcctcttggc, agccttcttcttcttcttccc
AT3G56790	3	21045058	Ts-1	A->T	PreStop	True	gacgcaagaaccttcattagg, agatcagaggaagagaatggg
AT3G57460	3	21274154	Bay-0	G->T	PreStop	True	tctggtaggttcgacaattcc, tgggtatgtcttgtggttagg
AT3G57680	3	21392662	C24	A->T	PreStop	True	agaactcttgtggaagcttgg, gtttatgaaaaggccaacacc

AT3G58200	3	21572035	Rrs-7	T->A	PreStop	True	cattttcagccttaagttcgg, gtagttggttcttgcaaatgg
AT3G58220	3	21576461	Nfa-8	C->A	PreStop	True	tgaaggaagaacttgaaaccc, gaagaggattacgaaaagagacc
AT3G58270	3	21588417	Rrs-7	T->A	PreStop	True	tgaaccatgctctaaacttgc, aatttctcctctcagcagtcc
AT3G58340	3	21601043	L*er*-1	C->A	PreStop	True	agcattctgatcaaaccagc, gatgtcattctgttccagtcg
AT3G58410	3	21616438	Tsu-1	G->C	PreStop	False	gaattctgatgcaatgtcagg, gttacagaaacacatcgctgg
AT3G58470	3	21638048	Bur-0	T->G	RevStop	True	acacctgagattggttaaggc, tcttctcacaggtacaatggg
AT3G58820	3	21764815	Bor-4	C->T	PreStop	True	caagcctcaagaccctaacc, aggtttttgaacaccggc
AT3G58910	3	21786294	Tamm-2	T->A	PreStop	True	tcacagcatagagagagcacc, gtattagattcggatgcgacc
AT3G59180	3	21893256	L*er*-1	C->G	PreStop	True	aatgttgggacgagagatacc, aggtgttgttgaggaagacg
AT3G59190	3	21896950	Van-0	G->T	PreStop	True	aggtttgctcatgactctcg, tgtgtctttgaagctgtaatcc
AT3G59270	3	21917826	Bur-0	A->C	PreStop	True	ttacttgaattctgcactcgg, acagtccaaacctagaaaccc
AT3G59300	3	21930894	Est-1	T->C	SD(non)	True	cctctagaagatttgaagccg, gatcaaaatgacacgcttacc
AT3G59550	3	22011472	Bur-0	G->T	RevStop	True	gaatgttcttcgagacactgg, aattcactccatcacaacacg
AT3G59750	3	22081758	Lov-5	T->A	PreStop	True	cgttcttgattctcattacgg, aagaagattttagcggtctgc
AT3G60590	3	22410020	Ts-1	G->C	Met	True	aagagtccgtctcaagaatcc, ggactgtgatgggtctttagg
AT3G60760	3	22469788	Lov-5	G->A	PreStop	True	cttcctttaagcattgatggc, gagagttatggcaggaaaacg
AT3G61350	3	22715303	L*er*-1	G->A	PreStop	True	gctaagctaacaatgtggtgc, gggttgttgtttatgaatcagg
AT3G61420	3	22739580	Rrs-7	C->A	SD	True	tggacatttctttttgacgc, atacgagaattgttgcagagg
AT3G61530	3	22782740	Est-1	A->T	PreStop	False	ttattaatcaagccaccaccc, gatagtttccgctgtgttgc
AT3G61940	3	22949227	Fei-0	A->T	PreStop	True	agttgttggagaaatccaagg, gaaaaccagagaaatgaaccc
AT3G62850	3	23249164	Van-0	C->A	PreStop	True	cagtagaaattccagagagatgg, gtggagggtttcaggagg
AT3G63320	3	23400920	L*er*-1	G->A	PreStop	True	gttgaaagtggttgagtctgc, tggctaatgacagctacttgg
AT3G63370	3	23416898	Nfa-8	A->T	PreStop	True	ggagatataggtatgcagggc, cttcacttgttcccttgtgg
AT3G63370	3	23416563	Tsu-1	C->T	PreStop	True	ttcagcagagaccttgtaagc, gccctgcatacctatatctcc
AT4G00070	4	29674	Van-0	A->G	RevStop	True	actcgtgagagagaatggtcc, cagctgtgtaacaatatgggc
AT4G00970	4	418971	Rrs-10	G->T	PreStop	True	ttaactgtattgctcaagccg, attacctggttttggatcagg
AT4G02190	4	968674	Lov-5	C->T	PreStop	True	caagaaaagatgtgtggaacg, gcagacttcttccatctctgg
AT4G02430	4	1071311	Rrs-7	T->G	RevStop	True	ttgcttctagttaagggtacacg, gcagaaaatcaaaccaacacaagc
AT4G02465	4	1081604	Nfa-8	T->G	PreStop	True	gtgctgctttgctaagtttgg, agaggctcggatcttttaagg
AT4G02660	4	1165347	Tamm-2	G->A	PreStop	True	catgatcaccatcttgtttcg, gcggcatagagacattgg

AT4G03090	4	1367831	Tsu-1	C->T	SD	True	agcttaccagataaattcccg, aatccttgatccctagttccc
AT4G03440	4	1527117	Rrs-7	G->A	SD(con)	True	gtcatctgttgacctctgtgc, agttaaggttacagggtcacg
AT4G03490	4	1552758	Bor-4	T->A	PreStop	True	taatgagagaaacttgggacg, actatccaaagcatgaacagc
AT4G03590	4	1602931	Sha	A->T	PreStop	True	attagctttttccatgtcgg, gcgtgaatctttttgtctttgg
AT4G03590	4	1600780	Est-1	A->G	RevStop	False	accaattcgttatggaatgc, gttggttgaagttgtgagagg
AT4G03600	4	1604160	Est-1	C->T	PreStop	True	cttgatttcccaaagaaggc, aaagaagcgtacacgacacg
AT4G03620	4	1608662	Bay-0	A->G	Met	True	aacttgacttgacgtttgagg, tggtctaagacataaaggagaagg
AT4G04110	4	1972590	Fei-0	T->C	Met	True	aaggtgaaatccaagtaagtgc, agtagtcaaaccctccactgc
AT4G04200	4	2027451	Sha	G->T	SD	True	aaacacctccttgatgaatcc, tacttgccaaagtcaagaagc
AT4G04390	4	2147220	Est-1	T->A	RevStop	True	tggaaaggtctctcttcatcc, agtataaccggcaacatacgg
AT4G04525	4	2251264	L*er*-1	G->A	PreStop	True	ccattagacccggtaactacg, ctttctcaattttccaacccc
AT4G04530	4	2252357	L*er*-1	G->A	PreStop	True	ctcatcctgagatcttcaacg, tcgtggaacagtaagatctgg
AT4G04545	4	2272364	Tsu-1	C->T	PreStop	True	ctatagaccaaaatggcatgg, tctccagccagaaaattgc
AT4G07480	4	4268967	Br-0	A->T	PreStop	True	attcaatggttacatccagcc, ttcctttgattctcttggagc
AT4G08013	4	4835909	Sha	C->T	PreStop	False	ccgcaataactattccagagc, ccaccacaatctaacacaagc
AT4G08013	4	4835937	Nfa-8	G->C	RevStop	True	ccgcaataactattccagagc, ccaccacaatctaacacaagc
AT4G08098	4	5006906	Bay-0	C->T	PreStop	True	aagatttcgtggataaggtcg, atttttggaagaggctcagg
AT4G08098	4	5006870	Tamm-2	G->T	PreStop	True	aagatttcgtggataaggtcg, atttttggaagaggctcagg
AT4G08130	4	5094315	Lov-5	T->C	SD(non)	True	ctcacaagtctcagttccagc, atttggtactggtgtcaatcg
AT4G08340	4	5267741	Bay-0	C->T	SD	True	cgttggaaaaggtctcacc, cggtcgtttcaattgtgc
AT4G08430	4	5347959	Bur-0	G->A	PreStop	True	tttttggcaagtcaatgtcc, gactccaagaagcgattgg
AT4G08560	4	5453057	Tamm-2	G->A	PreStop	True	aggtgacaagtgctctctcg, acaacaccaacaccaacacc
AT4G09060	4	5797815	Fei-0	C->T	PreStop	True	actgttcttgtagacgcaacg, aaagaagtaaacactgcgaagg
AT4G09360	4	5942063	Br-0	C->A	PreStop	True	agaagggaagatacacatcgg, tcagtcagctctacaaatgcc
AT4G09490	4	6015824	Fei-0	G->T	PreStop	True	tacacgattctcacttcgtgg, gagaagaagtcacagcagacg
AT4G09790	4	6164551	Bur-0	C->T	PreStop	True	tagtcttgtccttggtcaggc, aaggaaattggtaaacctaccc
AT4G09920	4	6225313	Br-0	T->A	PreStop	True	gctcgtgatcttgaaactcg, cgcgagatcaaagtcttgc
AT4G09965	4	6245264	Tsu-1	A->C	PreStop	True	cctggtagaagtagagacggc, tcagcgttaggagacagagg
AT4G10040	4	6278185	Nfa-8	G->C	SA	True	ttaagctggtgcatcataacc, caactagcctcttcaacaacg
AT4G10620	4	6566396	Van-0	T->C	RevStop	True	acgcgattttaggtgtatgg, gctattgtggtgaaacagagc

AT4G10740	4	6617850	Tamm-2	G->T	PreStop	True	ctaaacccctagccttaaacg, tcaagtctttcattgtgaggg
AT4G11040	4	6745840	Nfa-8	C->A	PreStop	True	agatgttgggacatcagaagg, tgatggatatgacagagaggc
AT4G12350	4	7326127	Rrs-10	A->C	Met	True	caagactttgacatctccacc, ctagaggaggttcatcgttgg
AT4G13730	4	7973640	Nfa-8	T->C	RevStop	True	acaaccaaaactgtaccatcg, gcccgttaactcaattctgc
AT4G14630	4	8393136	Tamm-2	G->A	SA	True	aaagtctctctttttcttgcgg, gtgtgaggtgggttctgacc
AT4G14820	4	8507866	Bay-0	G->A	PreStop	True	agacagatgctacgaaggagg, agaaggatttggttctatggc
AT4G14905	4	8527401	Bur-0	C->T	PreStop	True	caataaaagccgtacaacacg, ccatagattagttccggttcc
AT4G16095	4	9104946	Tamm-2	C->G	PreStop	True	tttcattgtacgaggtcatgc, cacccattcactattccttcc
AT4G16810	4	9461030	Fei-0	A->G	SD(non)	True	cttgagaatgcttcacatgc, ttatgtctgaccgagatagcg
AT4G16845	4	9478398	Bor-4	G->T	SA	True	agaggtggcagaaataacacc, aaaccttctagcctctgatcg
AT4G17280	4	9679322	Bay-0	G->T	PreStop	True	ggatgattatcgtgtagccg, aatattgaatggagtgagctgg
AT4G17565	4	9782817	Lov-5	T->A	PreStop	True	aatcagagaagagcatggagg, taccagtatgcttcctatggc
AT4G17860	4	9924929	Fei-0	C->A	PreStop	True	cacttcgtatttcaaactcgc, tccataagacaatctaccccg
AT4G17990	4	9985308	Bay-0	G->T	PreStop	True	agcaaaagtgccaatcttacc, cactgaagtcttccttaacgc
AT4G18330	4	10126676	Van-0	T->C	SD(non)	True	tgggcctagttaatcataagg, aaatgacttcaggaaacagagg
AT4G18720	4	10301237	Got-7	T->A	RevStop	True	cttatgcagccattaattccc, cttcaactatgaccttgtggc
AT4G18840	4	10338799	Rrs-10	C->T	PreStop	True	gatgtaacttcatgcttcttgg, aagcagaagaacttgtgaacg
AT4G19000	4	10406036	L*er*-1	C->T	PreStop	True	aagaggttcaagagatgtggg, actcaccttgaagaggtttcc
AT4G19030	4	10422493	Bor-4	T->A	SA	False	aaccgctctattatcggtagc, gtattgcacacgagactttgg
AT4G19080	4	10449449	Tsu-1	C->T	PreStop	True	acatgcatcttctgaaacagc, agatagtgttcacatcccgc
AT4G19360	4	10565417	Bor-4	A->C	RevStop	True	ttgaggcaatgactaagaacg, gagattcacaggtcagtaaggc
AT4G19470	4	10613801	Van-0	C->G	SD	True	agtccttgaaagagacaaccg, cagtacttggatgttcatggc
AT4G19560	4	10663786	L*er*-1	G->T	PreStop	True	aggtgctattctccttgctgc, cactttctgagttcacctctcc
AT4G19650	4	10693616	Br-0	G->A	SD	True	tctctttgttgttaggttgcg, acggttcaaactcgtttatcg
AT4G19730	4	10733907	Bor-4	G->T	PreStop	True	tgtctattaaacaccccatgc, cctggactctccaaaacg
AT4G19730	4	10734131	Tsu-1	C->A	PreStop	True	aataatgtccgacaaccttgg, atgagagaaagcttggattgg
AT4G19925	4	10800271	Rrs-7	T->A	PreStop	True	ttaataccctgaatgatcgcc, gatcaatgggaaaggagagg
AT4G20920	4	11195881	Sha	G->A	PreStop	True	taacatatttgcagtgggtcc, gagattctctgcatgttctgc
AT4G21230	4	11319533	Bur-0	C->T	PreStop	True	aagaaatcgaaaacgacgc, ttctgtagtggatatatggctcc
AT4G21840	4	11588188	Tamm-2	C->T	Met	True	tgcctcgtttactaaatcacc, cacattcaaaagtccacaagc

AT4G22110	4	11713412	Rrs-7	A->T	PreStop	True	aagtaacctgtggaaacaccg, ctaggattctaggacacgaagc
AT4G22250	4	11768340	C24	A->C	Met	True	tgtaactcttcgatttcgtcg, gggaaatcaatattcggaagc
AT4G22300	4	11789609	Bay-0	A->G	SD(non)	True	catgttcacactcaagattgc, gaatggtactctttgaagctgg
AT4G22730	4	11943013	Ler-1	A->T	PreStop	True	gaagacattgaatcagcaacc, gcttcttctgttgatcacttcc
AT4G23070	4	12090718	Tamm-2	T->A	PreStop	True	atcagtggattgagctttgc, ccatttcttcttctcatgttgg
AT4G23130	4	12119124	L*er*-1	C->G	SD	True	ccttgtaaacttgtccaaatcc, agagggaactgttcttcctcc
AT4G23200	4	12145928	Sha	G->C	PreStop	True	gatatggttagatccacgagc, tctagatgccgatatgatccc
AT4G23300	4	12183250	Van-0	A->T	PreStop	True	gttccagcagttgtcatcg, gtgtcttgtcctgatactttcg
AT4G23320	4	12190997	Lov-5	C->A	PreStop	True	gataatcgccacaatagctcc, aagctaacttgacggatttgg
AT4G23320	4	12190099	Tamm-2	C->T	SA	True	ttccaactactcttgctgtgg, aatcttgttaagcttctcggg
AT4G23410	4	12224954	Est-1	A->T	SA	True	atttggaaacagtctggatgc, taataccgatctctcccatgc
AT4G23420	4	12228567	Nfa-8	C->T	PreStop	True	ctagctttacggattgatttgg, aggctgaaattggataaacc
AT4G23520	4	12274940	Nfa-8	T->A	SA	True	gcatggtcaagatttgttcc, atcaggactaatggacacagc
AT4G23970	4	12445402	Sha	G->A	PreStop	True	tacagaggaaacaaatggtgg, aattgacctatgtgatggagc
AT4G24460	4	12644479	Bor-4	A->T	PreStop	True	gatctggtgcagatactacgc, tgaataacagaggaaagcagc
AT4G24600	4	12700410	Tsu-1	C->G	PreStop	True	cttctgtgagaactgctgacc, ggtaccacatctttctttagtgc
AT4G24700	4	12744818	Bay-0	T->A	RevStop	True	agagctttccttgaaacaacg, gatccctacgagattctttcg
AT4G24730	4	12753655	Bur-0	T->A	SA	True	aagccatcaacaatgtcacc, aaacaattctgacgaatggg
AT4G24980	4	12847593	Tamm-2	G->A	PreStop	True	attcaacttaccccagagacg, agtggctattcacagtcatcg
AT4G25160	4	12903372	Bor-4	A->G	RevStop	True	ggagaaatgaggattctcgg, aaaatgtgtgagcttgtgtgg
AT4G25380	4	12975957	Bor-4	G->A	PreStop	True	tagggtttgatcagtgagtcg, tgccctcctacgaatagtcc
AT4G25810	4	13129377	L*er*-1	C->T	PreStop	True	gttcttgggaaacctaagtgg, gttagggcatgaagactggc
AT4G25840	4	13140317	Lov-5	T->A	SD	False	gatcactgagagtgcaaatgg, ctcatcagctaagagaacttgg
AT4G26030	4	13206206	Nfa-8	T->G	RevStop	True	gttggatatatagcctgccg, ttctaggtgatttcccaatcc
AT4G26260	4	13297949	Rrs-7	T->A	Met	True	tctccaaatattaaggagggg, atattgtccccactcttctcg
AT4G27530	4	13752646	Br-0	A->G	Met	True	attttgaacatatgggctgg, cagctcaacgatttttgtatgc
AT4G27930	4	13902982	Sha	C->T	Met	True	tgatgttgaatggctctatgc, aaatcccaacacaaacctcc
AT4G27960	4	13917347	Bor-4	A->C	SD	True	tctgcagatccttcaattcc, aggtacaacgctgagttaggc
AT4G29200	4	14398921	Got-7	C->T	PreStop	True	gccattgataagaggagatcg, acacaaatcaaggaaggaagg
AT4G29550	4	14502853	Tsu-1	C->A	PreStop	True	tgtctgacgctgtacataacg, catcaactcaaggtttgagcc

AT4G31350	4	15210419	L*er*-1	G->A	PreStop	True	aaaagcaagagctatcttcgg, aagtatagagcagctcgcagg
AT4G31400	4	15239022	Lov-5	T->A	PreStop	True	tggatgtctagttgctgaacc, agatcttcctatggagcttgg
AT4G31520	4	15280936	L*er*-1	C->A	PreStop	True	ctaaacatcggaaacaggacc, agcaagaagatgaaaccaagg
AT4G31710	4	15351118	L*er*-1	C->T	PreStop	True	ttgaacatctacggattgagg, ctgagggaaagtaccaacagg
AT4G31760	4	15369730	Rrs-7	T->C	Met	True	aagaaaagacgaaggagtttcc, tggattcaactaaacacgacc
AT4G32520	4	15692012	Bor-4	T->C	SA	True	ttacctttaccaggcagtcc, gttcaaggatctgtctttccc
AT4G32990	4	15921097	L*er*-1	T->G	PreStop	True	aaatacaaccagaggaggacg, ttcaccaacctacaagtgacc
AT4G33130	4	15979768	Bay-0	C->T	SA	True	ctctctcttttcctcgtcacc, gcagcagcagttacaagagg
AT4G33290	4	16050806	Rrs-10	G->A	PreStop	True	acgaggagaagaagaaagtcg, gattctgttccaaattcctcg
AT4G34460	4	16477629	Lov-5	C->T	PreStop	True	aatcactctcctgtgtcctcc, accaactccaggtctatcagc
AT4G35820	4	16971231	Lov-5	A->T	PreStop	True	gacaggtgtcactctattacttgc, cgtaacgaagcagcaacc
AT4G37590	4	17663202	C24	T->A	PreStop	True	ggaagaacaaaaccacaaagg, aagtaagagccaacaacaacg
AT4G38510	4	18013206	Fei-0	G->A	SD(con)	True	gaacagagcatgcaaatatcg, tgtccttctcctttgatttgg
AT5G01050	5	18549	C24	G->A	PreStop	True	cctaatggtaaatgtgcatcc, tgacatggagatgatgtttcc
AT5G01150	5	53130	Bor-4	C->T	PreStop	True	atgaagcagaacctgaaaagg, tcttcaagaacccctgagc
AT5G01760	5	294122	C24	T->C	RevStop	True	gttatcctcagccacaatgg, aactaattgaggaggctttgc
AT5G05280	5	1565735	Br-0	G->T	PreStop	True	ccaaaccagaaagaagaaaacc, acatggtgatcatactagccg
AT5G06440	5	1966379	Fei-0	A->G	Met	True	tgtagagagttgattcccacg, gatgtatccaagttacttagcaagc
AT5G10140	5	3173827	Bur-0	C->G	SA	True	acgctcgcccttatcagc, gtggctcagttccaactcc
AT5G10250	5	3218326	Br-0	A->T	PreStop	True	acattagcatcttccaaagcc, taaggagatgctgtgacttgc
AT5G10800	5	3415524	Fei-0	C->A	SA	True	agactcttttccatgctctgg, ttatttctcctgaggacgagc
AT5G10850	5	3428605	Fei-0	G->T	PreStop	True	tgataacaattggcagtgagg, gatttcggtacaaagtttccg
AT5G14970	5	4847426	C24	C->G	PreStop	False	caaaatctttggcttagtggg, gatcacagcgaaacctcg
AT5G16330	5	5346556	Sha	G->C	PreStop	True	ttcccaacacatagtcttttcc, cattcatttcacttgtggagg
AT5G17250	5	5670087	Tamm-2	C->A	SA	True	agcattgatccttgcttcc, gctagggatagctccagacc
AT5G18710	5	6242025	Bur-0	T->C	SD(non)	True	tggtgagaagagaaagagaagc, aacctaccaacagaaacaggc
AT5G19720	5	6668035	Bur-0	C->A	PreStop	True	tgggttgttgatctaggttcc, cttagttggctcttttcctgg
AT5G20220	5	6825736	Bor-4	C->G	PreStop	True	cactaaaggcatttccactagg, tacttgcacgtaaacgaatgc
AT5G20230	5	6827408	Bur-0	T->G	RevStop	True	tatgctaaacaccactggacc, cccactctttattttgcaacc
AT5G20430	5	6904910	Bur-0	G->A	PreStop	True	ttatgaaggctggaaggtagc, gaagcaagtgttgagatgagc

AT5G22160	5	7349211	Bay-0	G->T	PreStop	True	caaaccagatgtccttttcg, gattgtgtggctttgtaaacg
AT5G22450	5	7442930	Bur-0	A->T	PreStop	True	tactaggaacaccgagaaccc, gtctttgattcatggcttgg
AT5G23580	5	7952287	Bur-0	G->T	PreStop	True	aactcacctccttcacaaagc, tctctcttcaatgccttctcc
AT5G25600	5	8913231	C24	G->A	PreStop	True	ctgaggagcaacagtcatagc, tgatcactgtccttatctggc
AT5G25920	5	9044808	Bur-0	T->A	PreStop	True	acttcattgtgtttcaccacg, accaaccctcagtctctaagc
AT5G27300	5	9621827	C24	C->T	PreStop	True	gaaagctgggaagttgatacg, catcaatctcaccactaagcg
AT5G27800	5	9842849	Sha	T->A	SD	True	tttttggttctatctcggagc, acccctagccttactctctcc
AT5G28190	5	10168563	Bur-0	C->T	PreStop	True	gctttctaatcagagcacacg, gaaagttaagcttcgtagtggc
AT5G28270	5	10258538	Br-0	G->A	PreStop	True	gtaccagacgtacctgattgg, gaagcgtttgataagtatgcg
AT5G28295	5	10283287	Bay-0	T->C	RevStop	True	cgactctaataacgaaaagcc, tagaggttgccgagattttgc
AT5G28420	5	10364916	Tsu-1	C->T	PreStop	True	cagatcttccaaaacgaaagg, tgagcaagtgaaatgttctcc
AT5G28820	5	10832098	C24	C->T	PreStop	True	ctacaacgaagaaattcacgg, ttcgacttccttttcttcagc
AT5G31412	5	11560213	C24	G->A	PreStop	True	cttaggcagcttagaaatggc, gcttttaggatgttttgtgagg
AT5G32070	5	11466214	Lov-5	G->C	SA	True	aagtgctgatagcattgatcc, acaatgcaacatacagttggc
AT5G32613	5	12280596	Nfa-8	G->C	SD	True	ataaacttgccatcaaacgg, ccagaaccttaggagatgaagg
AT5G34860	5	13200455	Bor-4	G->T	PreStop	True	cctatgacaagtcaacaacgc, gaacataaccgagatccaagc
AT5G35120	5	13404084	Lov-5	A->G	RevStop	True	cacgatttaaaggaaaaccc, aaatcgagttatgaaggctcg
AT5G35600	5	13787997	Bur-0	G->A	PreStop	True	ttgtcaagttgttctccaacc, ataagtgactatggggaaggg
AT5G35604	5	13805530	C24	G->A	PreStop	True	gatagcctttggatcaactcc, caattatatcgtcttgcaggc
AT5G37120	5	14695016	Sha	G->T	PreStop	True	accgtcgattatagtgaaacg, ccacacctaacacattcatcc
AT5G37410	5	14854561	Fei-0	T->G	PreStop	False	tgttgtggagttgcataaagg, acacaaaatggcattgatcc
AT5G38840	5	15568799	Lov-5	C->T	PreStop	False	gaattctctaatctctctgcaacc, ctgcaaagggaacaaatacg
AT5G38900	5	15592187	Tamm-2	C->G	SA	True	cagatggatcaaggaaaaacg, agccacctgatattgaagagc
AT5G39100	5	15670660	Br-0	G->T	PreStop	True	ctaaacttggcctcaagatcc, tccagggttaaacactatggg
AT5G43240	5	17370940	C24	A->C	PreStop	True	tgtccctctcttcataaagcc, gagaggaatcaacacctgacc
AT5G44970	5	18173197	Bay-0	T->A	SD	True	tcggagagttcttctgattcc, agagaaggatttttggtctcg
AT5G45000	5	18182942	Bor-4	C->T	PreStop	True	cttcagaggagaggagctacg, tataatacctgtgtcctgccg
AT5G45180	5	18293062	Bor-4	C->A	PreStop	True	gctaatgcagactctaacgcc, tgcaaatccatattgttggg
AT5G45640	5	18525991	Br-0	G->C	PreStop	True	ggtttgatgaaagcaaccg, tgctcctctagtctacgctcc
AT5G46140	5	18722971	Sha	C->A	PreStop	True	ttacactcgtccccatagtacc, catagcgatgtcctcttgtcc

AT5G46875	5	19041773	Bor-4	G->A	SA	True	attcatattcttcgggactgc, aacatcgaaaccatcacacc
AT5G46980	5	19083112	Fei-0	G->T	PreStop	True	ctcatagccctcagaataaagc, gaagtaagcggagtggtaacg
AT5G48375	5	19618671	Bor-4	T->G	SA	True	aggtccaaacgaacaataagg, ctaaatcggtccaagaaatcg
AT5G49050	5	19901197	Tamm-2	G->T	PreStop	True	aattgaactctctccactttgg, gtataatatcgaccgttcccg
AT5G49500	5	20095650	Tamm-2	T->A	PreStop	True	catcacgttgtgctctttagc, tgctgagttgtcatgtactcg
AT5G49840	5	20273906	Bur-0	T->G	SD	True	ttatccctcttaaggtttggg, agtggatcaagaagaaagtgg
AT5G51580	5	20970275	Bay-0	A->C	PreStop	True	tcgactatctcttccaaaccc, tcttcaactcgtctacgaagc
AT5G51795	5	21060630	Bur-0	C->A	PreStop	True	tgtattcgaaccacgatatgc, tttaggttgaagaaggttggg
AT5G52150	5	21207233	Br-0	A->T	PreStop	True	tagcattatggaacaacctcg, gtggcttcaaattctgtatgg
AT5G52290	5	21251797	Bur-0	A->T	PreStop	False	gcctgaatattttctgaaggg, cttcagaaggagacaaatggg
AT5G53010	5	21511567	Tamm-2	C->A	SD	True	agaagaagagtcagcattgagg, aaactcattgaaagttgccg
AT5G55200	5	22412645	Bur-0	G->C	SA	True	gaatgttcttcaaaggttcgg, cgcggagtagagaagtaaacc
AT5G56990	5	23079418	Bur-0	G->C	PreStop	True	ctagtctcgaaccgaaaatcc, taaggcagcattccttttcc
AT5G58180	5	23561564	Tamm-2	C->A	PreStop	True	cttgctcgttctcaagtgc, ctccatgtgtcaccataatcc
AT5G61180	5	24630130	Fei-0	C->G	PreStop	True	cactaatttagggtgtctccg, ttcgacagaactgatctaatgc
AT5G62120	5	24964585	Bor-4	C->A	SA	True	gtagccaatcatctggatcg, aagcaagaagaatccaagagg
AT5G62970	5	25290248	Sha	G->T	SA	True	tgcgcttatgaagggtatagg, aagttagatgaaagggcaagg
AT5G64060	5	25651055	C24	T->A	PreStop	True	cataacatagaaagctggctcc, tgattgcttctgaaactacagg
AT5G64910	5	25959990	Br-0	C->T	PreStop	True	tacatcccacttgaacaaagg, atacaccaaattctgcactcg
AT5G66830	5	26709698	Sha	C->G	PreStop	True	gccaattccatctacttctcc, catgattctcttgatcatccg
AT5G67050	5	26776710	Bur-0	A->G	RevStop	True	aaattactctcttcaacgccg, actatagtggctgagaagggc
AT5G67530	5	26958488	Bur-0	G->A	Met	True	ctccggctttttaagtaatcg, gattgtacacacaggatcttcg

Table S11. Overlaps of deletions and highly polymorphic regions to the coding portions of genes as ascertained by dideoxy sequencing of PRPs.

Notes:
[a] Coordinates for PRPs queried by dideoxy sequencing [Chromosome (Chr), "Start" and "End" refer to core PRP prediction].
[b] Deletions ≥ 50 bp that overlap coding sequences are listed with effects on gene models. When a deletion ≥ 50 bp was not observed for a given validation attempt, the number of polymorphic sites (SNPs, deletions < 50 bp, or insertions) within reads [polymorphism number (PMN)] is given. The length of a "Polymorphic region" corresponds to the extent of available sequence. The extreme nature of the polymorphism underlying some PRPs, as well as the lack of double stranded sequence, confounded alignment for some sequences, and PMN instances may be overestimated for some alignments. Exon annotations are for coding exons.
* Ambiguous/complex deletion alignments.

Gene	PRP coordinates[a]			Acces-sion	Description[b]	Primers used for validation (Forward, Reverse)
	Chr	Start	End			
At1g03710	1	923578	923957	Nfa-8	Deletion of 379 bp partially removing exons 1 and 2	tcgtaaacatgatcgctaacc, ttgatactccacaagctctcc
At1g09850	1	3202878	3204108	Sha	Deletion of 1214 bp removing exons 3-5, partially removing exon 2	agtaggcaatgtaggaatttgg, cagcgagaggtttaagaaagg
At1g14660	1	5032492	5032809	Got-7	Polymorphic region of 735 bp (PMN=84) overlapping exons 15-18	cagttcatgcatctgtctcg, gtactgccaatgttttgatgc
At1g17300	1	5927269	5927762	Cvi-0	Deletion of 69 bp partially removing exon 1*	agctttcaagaatcctaacgg, ggtgattgattagtccagtcg
At1g21160	1	7409259	7411105	Cvi-0	Deletions of 1068, 883 bp removing exons 2-8, partially removing exon 9	atccttcttatgcacaggtcc, ctaaaggtgatggggaaacc
At1g23840	1	8425307	8426870	Cvi-0	Deletion of 1714 bp partially removing exon 1	cctcagatgtgtaagcaatcg, gatttggtttcccacttatgc
At1g23850	1	8425307	8426870	Cvi-0	Deletion of 1714 bp partially removing exon 1	cctcagatgtgtaagcaatcg, gatttggtttcccacttatgc
At1g30010	1	10515290	10515623	Bor-4	Polymorphic region of 825 bp (PMN=53) overlapping exon 1	gaaggagttcaattctctcgc, caccttgctaataccaccg
At1g31390	1	11242500	11243551	L*er*-1	Deletions of 171, 76, 88, and 79 bp partially removing exons 3 and 4*	agcgagacttctgatcaaacc, ctctggtttccaaacaatgc
At1g31510	1	11277731	11279565	Fei-0	Deletions (total of ~2400 bp) partially removing exons 1-3*	acagtcttcaattacgttccg, ataaaccctatggatttgggg
At1g31520	1	11277731	11279565	Fei-0	Deletions (total of ~2400 bp) partially removing exons 1-3*	acagtcttcaattacgttccg, ataaaccctatggatttgggg
At1g31620	1	11317314	11317936	Ts-1	Deletion of 682 bp removing exons 3 and 4	tgattcgtgtcaaagatgtgg, gggtaatgtatttctgcgcc
At1g31835	1	11423544	11424109	Bay-0	Polymorphic region of 912 bp (PMN=129) overlapping exon 1	atgtccttcgatacaaggtgc, tgacctatccctaatcgaacc
At1g31840	1	11423544	11424109	Bay-0	Polymorphic region of 912 bp (PMN=129) overlapping exon 1	atgtccttcgatacaaggtgc, tgacctatccctaatcgaacc
At1g33530	1	12160237	12161867	Bor-4	Deletions (total of ~1750 bp) removing exon 2, partially removing exon 1*	aagagaaagaagtggcagtcg, gatatctttgatcccaccacc
At1g35750	1	13255191	13255492	Cvi-0	Polymorphic region of 967 bp (PMN=68) overlapping exon 2, 3, and 4	gacacttctcaatcacatggc, ggaattacatcgtccagaagc

At1g37020	1	14052136	14052779	Bor-4	Deletion of 648 bp partially removing exon 8	ctacgtttacgacagcattcc, tgtgacgattagtgagaagcg
At1g37080	1	14113646	14114374	Bor-4	Deletion of 790 bp removing exon 2, partially removing exon 1	tgcatatgcttgtcttctagc, aacaagtcaacatgaaacccc
At1g41810	1	15582516	15582919	Cvi-0	Deletion of 391 bp removing exon 2, partially removing exons 1 and 3	cttaatatcgaaagggttggc, gtatctttctctaccgagccc
At1g44770	1	16910510	16911111	Fei-0	Polymorphic region of 1016 bp (PMN=102) overlapping exons 3-6	caaagagccctaagaacaacc, caatttccattcaaggaacc
At1g47940	1	17670625	17672084	L*er*-1	Deletion of 636 bp removing exon 1	acaccaatccaaactgaatcc, gcatttcagagacaaaaacacc
At1g52990	1	19746258	19746738	Tamm-2	Deletion of 547 bp removing exon 2	actgaagacaatgattcggg, tactgacgataacgtgcttgg
At1g57906	1	21439858	21440843	Rrs-7	Deletion of 779 bp removing exon 3	caattcaaccaattcgaagc, tttgctagaggagtgaatccc
At1g59620	1	21909030	21909683	Lov-5	Deletion of 338 bp partially removing exon 5	aaggattagttttggactgcg, ttcaagaaaaggaccatgagg
At1g60540	1	22307040	22307624	Bay-0	Deletions of 433, 142 bp partially removing exon 2	ctccttgatgaactcaactgg, gagatatcccacattcaaaacg
At1g61940	1	22901800	22902367	Fei-0	Deletions of 131, 358 bp removing exon 1, partially removing exon 2	gctctgttcttcccatctagg, caagtggctgtctttaattagc
At1g66880	1	24949952	24951325	Tamm-2	Deletion of 1374 bp removing exon 1	ggtgtttcgattgtgaacg, aaggaagtatgtatgtggcacc
At1g67455	1	25271309	25272235	Ts-1	Deletions (total of ~1200bp) removing exons 1 and 2*	tcgaggaaaagaaaagatcg, aaatggtagaggaagactcgg
At1g69730	1	26233622	26233986	Ts-1	Polymorphic region of 805 bp (PMN=120) overlapping exons 1-3	tacactgtaccttgaccaccc, gaaaacaccataacgagaggg
At1g74170	1	27895790	27896506	Bay-0	Deletions of 141, 70, and 608 bp partially removing exon 7	ttccactccattatctgttgg, gaaatctgccatcttctctgg
At1g76960	1	28925502	28925944	Cvi-0	Deletion of 392 bp removing exon 1, partially removing exon 2	ttgattggtgaccatttgc, tgcagtctaagagagttgtttgg
At2g04420	2	1535252	1536076	Ts-1	Deletion of 997 bp removing entire gene	gagggagagagatcgtactgc, ccgatatttttgattacccg
At2g05900	2	2257456	2257978	Got-7	Polymorphic region of 901 bp (PMN=240) overlapping exon 1	ggacatgtatcatggacaacc, gacctaatcaaacacgagaagc
At2g05915	2	2263749	2264199	Bay-0	Deletion of 425 bp removing exon 1, partially removing exon 2	agctgcactaaccaaggtagc, tctctctttctctctggcacc
At2g16870	2	7315756	7316223	Sha	Deletion of 497 bp partially removing exon 4	ataatcctgttatcccttggc, ggaactatgtgattgcaggg
At2g19550	2	8471718	8472705	L*er*-1	Deletion of 1031 bp partially removing exon 1	ctaccagcctgaaagacaagg, tcctttccctaaactaaacgg
At2g26610	2	11329126	11330682	Cvi-0	Deletion of 1544 bp removing exons 12-17, partially removing exon 11	tcatcttcgtttaacacctgc, aggttgtgttcaatgttcacg
At2g27600	2	11788947	11789437	Cvi-0	Polymorphic region of 885 bp (PMN=71) overlapping exon 2	tgcgattgttagagagaaacc, ttcactctcgtttccttctcc
At2g27760	2	11832733	11833044	Tsu-1	Polymorphic region of 876 bp (PMN=132) overlapping exons 3-6	ttttcgagacttcactgttcc, tctacctctgcagtctttccc
At2g35075	2	14795706	14796442	Bay-0	Deletion of 486 bp removing exon 11, partially removing exon 10	aaatccgtattaggttgcagg, ttcgtacgtttgatcttctcg
At2g35080	2	14795706	14796442	Bay-0	Deletion of 233 bp partially removing exon 6	aaatccgtattaggttgcagg, ttcgtacgtttgatcttctcg
At2g42470	2	17686860	17687919	Rrs-7	Deletion of 1047 bp removing exon 10, partially removing exon 9	gttaaccaagaaaaatctcccc, caaagaacttcatcgaacacc
At3g04660	3	1265724	1266292	Br-0	Deletion of 244 bp partially removing exon 1	tcatcattctggatctcaagg, tgtaacttacgaaggcgagc
At3g05450	3	1576194	1577106	Sha	Deletion of 705 bp removing exon 1	ccctaaacaaaccaaagatacg, acgttgaattgaggaaactcc
At3g09160	3	2805710	2806168	L*er*-1	Deletion of 440 bp removing exons 4 and 5	aataagaaagagcagcatgagg, aactcaatggaagtgacatgg
At3g11405	3	3580759	3581471	Tamm-2	Deletion of 674 bp removing entire gene	ccagtttgtttgttggtttgg, ggtcgattttggtcctagc

At3g14460	3	4853787	4864328	Cvi-0	Deletion of 10536 bp partially removing exon 1	ctccagaaactgctcttaggg, tcctgaagttacaagcctcg
At3g14470	3	4853787	4864328	Cvi-0	Deletion of 10536 bp removing entire gene	ctccagaaactgctcttaggg, tcctgaagttacaagcctcg
At3g14480	3	4853787	4864328	Cvi-0	Deletion of 10536 bp removing entire gene	ctccagaaactgctcttaggg, tcctgaagttacaagcctcg
At3g14490	3	4853787	4864328	Cvi-0	Deletion of 10536 bp removing exons 6 and 7	ctccagaaactgctcttaggg, tcctgaagttacaagcctcg
At3g16030	3	5440255	5440732	Est-1	Polymorphic region of 904 bp (PMN=184) overlapping exon 2	atcttcagctccaagagatgg, gtgatgccacaacaactaacc
At3g16520	3	5618702	5619096	Rrs-10	Deletion of 434 bp removing exon 3	agacttctttgcactgacgc, aaactggttcgtctcataccg
At3g17200	3	5870017	5870834	Tamm-2	Deletion of 959 bp partially removing exon 1	tagtgtttacacatgcgttcg, atgtaacgtacccgaacttgg
At3g18270	3	6262671	6263046	Cvi-0	Polymorphic region of 749 bp (PMN=75) overlapping exons 2 and 3	aggatcagataacggctatgg, aaacctctagcgctcaatacc
At3g18485	3	6342407	6343398	L*er*-1	Deletion of 983 bp removing exon 1, partially removing exon 2	gctggtaaccgaactcatagg, gattttgatcacttgtgtcacg
At3g19040	3	6571007	6571502	Got-7	Deletion of 484 bp removing exon 9, partially removing exon 8	tgagaacagattgatcatgcc, gacaacaggtttgtgttttgc
At3g21080	3	7385876	7391501	Fei-0	Deletions of 2729, 1768 bp removing entire gene	gcgtacttttgtgagagactacc, acagacttctcttcttcgtgg
At3g21960	3	7737061	7737635	Rrs-7	Deletions of 135, 119, 256 bp partially removing exon 1	gtgtgtagttgtggttatgcg, ctacttgctttcgcctttagc
At3g22080	3	7779531	7780311	Van-0	Deletion of 760 bp removing exon 5	gcttgaacacatcattgaacc, gtgaaacactcaatgcagagg
At3g22860	3	8091410	8092722	Nfa-8	Deletion of 1285 bp partially removing exon 1	accatatcgagtatgctgtcg, taaacgttttgaggagatggc
At3g23960	3	8658116	8658623	Bay-0	Deletion of 451 bp partially removing exon 1	gtgaaatccatagcaacatgc, ttaaacgcttctaatcctcgg
At3g25080	3	9136744	9137391	Ts-1	Deletion of 638 bp removing exon 1	tcccgcgatttattagtgg, gaccagaatcactagcttccc
At3g27590	3	10222724	10223082	Bur-0	Deletion of 391 bp removing exon 2	ttgcgcttctaatatcctcg, atgtggctttgataattggc
At3g27600	3	10225406	10226165	Bur-0	Deletion of 199 bp removing exon 3	caaagtgaggatttctcctgc, aatgaccaacgtcaaagatcc
At3g27910	3	10359054	10359510	Bur-0	Deletion of 427 bp partially removing exon 1	ataacgcacgaaccaactacc, ctgggactagaaatctttggc
At3g28140	3	10469935	10471687	Sha	Polymorphic region of 206 bp (PMN=7) overlapping exon 1	atagcagtgattatgggagcg, ggttcaagctttgttgaatcg
At3g28260	3	10535663	10536057	Bur-0	Deletion of 473 bp removing exon 1	aagcatacgaatgatactgcg, acattctccaaaacgctatcc
At3g28680	3	10750191	10751192	C24	Deletions of 779, 240, and 141 bp removing exons 2-4, partially removing exon 1	gaagatgcaagctaagactcg, gcattctcaatagcgttttgg
At3g28880	3	10894878	10896159	Bay-0	Deletion of 1279 bp removing exons 5-7, partially removing exons 4, 8	gaatcgattgaagactgatcg, aggaactgataaggcttttgg
At3g29250	3	11198704	11199088	Bay-0	Deletion of 320 bp removing exon 4, partially removing exon 3	tccgatgtgcaacaatatacc, gatgcagaatagttgaattgcc
At3g29330	3	11259236	11260351	Ts-1	Deletions (total of 1300 bp) partially removing exons 1 and 2*	gtaatttacaagggccttcgc, gtgttacagacatgaaacagaagg
At3g32904	3	13455322	13455951	Bay-0	Deletion of 577 bp removing exon 3, partially removing exon 4	gcatttggtggtgaactcc, caacatctaactcatggtggc
At3g32930	3	13494377	13494802	Bay-0	Polymorphic region of 912 bp (PMN=93) overlapping exon 4	aaggatgcagctgataagagg, gcaccatttcatggtaacg
At3g32940	3	13494377	13494802	Bay-0	Polymorphic region of 912 bp (PMN=93) overlapping exon 8	aaggatgcagctgataagagg, gcaccatttcatggtaacg
At3g33293	3	14047957	14048540	Bur-0	Deletion of 576 bp removing exon 2, partially removing exon 1	gcaattattacctcaaaggcg, accaaagatacgacaagctcc

At3g42200	3	14382378	14387370	Cvi-0	Deletion of 4984 bp removing entire gene	cacctaaattttctcgctgc, aacacaaatgaccctaggagg
At3g42820	3	14932639	14933260	Bay-0	Deletions (total of ~570 bp) removing exons 19-21*	ttcagtcctcaacaaatgacg, tcatgtttaccgtatatgcacc
At3g42910	3	14987933	14988630	Bay-0	Deletions of 499, 221 bp partially removing exons 1-3	cttgaaaagattgaccaaccg, gatcctaaatgcctaagtgtgc
At3g43760	3	15659400	15659750	Tamm-2	Deletion of 432 bp partially removing exons 2 and 3	catgccagatagtgtaaacgg, cccttcttgaagttgatttgg
At3g44070	3	15839540	15839850	Cvi-0	Deletions of 248, 244 bp partially removing exon 1 and 2	accaaatacaaagatggaggg, ttagcaaccatgttcaaggc
At3g44290	3	15983784	15984297	L*er*-1	Deletion of 482 bp removing exon 4	tggacttacatgtgtcttccg, gttatttgccgacttaggagg
At3g44805	3	16360847	16361700	Bay-0	Deletion of 843 bp removing exon 1, partially removing exon 2	aaacttcatcactcattagggg, tgcatagtcatcaagaatgagg
At3g45010	3	16477633	16478011	Cvi-0	Polymorphic region of 688 bp (PMN=78) overlapping exon 1	aaatctctgacagaaaagccc, aactgaaaccagttcctaccg
At3g45820	3	16850777	16851663	Bay-0	Deletion of 729 bp removing exons 3 and 4	ggcttctctgacctacaatgg, ggaaaccctagagaaggaacc
At3g45940	3	16900051	16900659	Br-0	Polymorphic region of 1076 bp (PMN=108) overlapping exon 1 and 2	attacattcggtggttgttcc, aggaactgtaaggaatttgcg
At3g45950	3	16900051	16900659	Br-0	Polymorphic region of 1076 bp (PMN=108) overlapping exon 1	attacattcggtggttgttcc, aggaactgtaaggaatttgcg
At3g46530	3	17142295	17142673	Br-0	Polymorphic region of 828 bp (PMN=104) overlapping exon 1	gttattccaaccagtgtcacg, ccagaggactatgagattgacc
At3g46800	3	17246675	17246995	Fei-0	Deletion of 375 bp partially removing exon 1	atcagattctcatcgaccacc, caatatgcgacttcaatgtgc
At3g50810	3	18898875	18899375	L*er*-1	Deletion of 609 bp removing exon 4	tatgtcgtttcttcatcaccg, gtggtagcttcttttgaaaacc
At3g55590	3	20628684	20629118	Nfa-8	Deletions of 71, 297 bp removing exons 2 and 3	ttgtcttaatctggacttgcc, agcattgcatagagcttttcc
At4g07380	4	4191798	4192563	Fei-0	Deletion of 89 bp partially removing exon 2*	tttcctgcatgttcacttagg, tacggatttattgtagcagcg
At4g07510	4	4312873	4313284	L*er*-1	Deletions of 267, 89, 122 bp removing exons 3-5*	ctcatctcccgtgataggc, gcttggaagaagcagttatgg
At4g13130	4	7646887	7647930	Cvi-0	Deletions of 269, 468 bp partially removing exon 1	tgtcgaagatagtttcgatgg, agaatgtcttgggtgaagagg
At4g14600	4	8377276	8377626	Cvi-0	Polymorphic region of 920 bp (PMN=88) overlapping exon 3	tgtctgtgttttcattacagcc, gctctgatcaatgtcatttgc
At4g17990	4	9987783	9990887	Nfa-8	Deletion of 710 bp removing exon 1	ttatcaatctcgtacaccagc, tatacatgggattcgtttggg
At4g18000	4	9987783	9990887	Nfa-8	Deletion of 2362 bp removing entire gene	ttatcaatctcgtacaccagc, tatacatgggattcgtttggg
At4g18330	4	10127332	10127758	Bay-0	Deletion of 740 bp removing exons 4-5, partially removing exon 6	aagtgtgaggatgacaagtgc, aactgaagctcgttctgttcc
At4g19470	4	10613063	10613799	Rrs-10	Deletions of 436, 244 bp partially removing exon 3	gttaacatatgcgaggtggc, catttctcttccagtgcttcc
At4g19630	4	10684622	10686534	Van-0	Deletion of 1953 bp partially removing exon 1	ctcttgtaaagccctaccacc, ttgagagagcgttattagttgc
At4g23240	4	12161883	12163505	Cvi-0	Deletion of 1275 bp partially removing exon 1*	cacatccaacgtatagagcc, gttccttctctcgttccg
At4g23250	4	12161883	12163505	Cvi-0	Deletion of 1275 bp removing exons 10, 11, and 12*	cacatccaacgtatagagcc, gttccttctctcgttccg
At4g23510	4	12267856	12270088	Sha	Deletion of 2238 bp removing exons 2 and 3, partially removing exon 1	ctgaaacattgaagaagcagc, atgttttcatggacgagtacg
At4g24410	4	12623782	12624745	Lov-5	Deletion of 85 bp partially removing exon 1	actgctctcttctccatctcg, ggtgatactcagcatctcagc
At4g26280	4	13305089	13305394	Van-0	Deletion of 298 bp partially removing exon 2	ccggagttagatgattcttcc, cttcagagaaagggtacctcg

At4g26410	4	13347009	13347486	Bur-0	Polymorphic region of 828 bp (PMN=78) overlapping exon 1	gttgaaagaaggagaaaaggc, ggataacaaaagcagcagagg
At4g27430	4	13721371	13721835	Sha	Deletion of 442 bp partially removing exon 8	aagggaacaactatgacgagg, actcgatttcttttcctgagc
At4g28350	4	14025921	14026782	Sha	Deletions (total of ~1000 bp) partially removing exon 1*	tcttgtagagttatccttgtgatcc, ctgatgtcgttgaactcttgg
At4g29090	4	14333444	14335295	Cvi-0	Deletion of 1864 bp removing entire gene	gcgaatcttaatcctttctcg, acttttgtggatggtaacacg
At4g31740	4	15363190	15363817	Rrs-7	Deletion of 604 bp partially removing exon 1	acatttgagatagtggagggg, gaggtatgaatcggtttctgg
At4g32200	4	15550801	15551253	Nfa-8	Deletion of 467 bp partially removing exons 9 and 10	ggatacaaaaggaagacctgc, tcgttgagcaaatatctcagg
At4g33810	4	16213505	16214368	Tamm-2	Deletions of 76, 308, 472 bp partially removing exons 4 and 5	ccggtaggaagaaaacacg, gtctggccagttgttttgg
At4g34780	4	16592337	16592858	Lov-5	Deletion of 682 bp removing entire gene	tatgcctcaaacatctcctgc, atctgataccatttttgcgg
At4g37620	4	17674162	17674595	L*er*-1	Deletion of 468 bp removing entire gene	caacaatcatgcctataacagc, atgcgactatctccattctcc
At5g02660	5	602052	602625	L*er*-1	Deletion of 564 bp removing exon 2, parially removing exon 1	aacaactgctactaggggagc, gaagaatctcaacagtggaagc
At5g03500	5	875976	876569	Bor-4	Deletions of 333, 243 bp removing exon 6	ccaaaattaagactcttcccg, tcttgttgtataggcgagagc
At5g04400	5	1241656	1242603	Tsu-1	Deletion of 1074 bp removing exons 1 and 2, partially removing exon 3	gcacaaaacgacaagaaacg, catctaacataccgttgagcg
At5g09910	5	3092826	3093306	Lov-5	Polymorphic region of 965 bp (PMN=88) overlapping exon 1	cagataattcgcagagttccc, aagagatgattttcccacacc
At5g15990	5	5217415	5221088	Bay-0	Deletion of 3917 bp removing gene	ggtttcttcgaaataagcgg, aagagcatatggaatggaagc
At5g17680	5	5825917	5826602	Br-0	Deletions of 390, 146, and 114 bp partially removing exon 4	ccttaatgtcaacgagttccc, tcgactgagaattcaaacagg
At5g17780	5	5867059	5867522	Cvi-0	Deletion of 417 bp partially removing exon 4	ttttcaatcactcacccacc, gcaagcatcataagatatggg
At5g18880	5	6300462	6301483	Nfa-8	Deletion of 1074 bp removing entire gene	ctgaattcaacatcttcaccg, gttcgtttgaagtaacaacgg
At5g25120	5	8664403	8667128	Lov-5	Polymorphic region of 487 bp (PMN=19) overlapping exon 2	gctcgttccacttctaattcc, ccataaaagttgaaactctccc
At5g25415	5	8836452	8838240	Est-1	Deletion of 1965 bp removing exons 6-8, partially removing exon 5	agctaaacgttggacgatacc, gtgaatacacaaaacagttggc
At5g25920	5	9044169	9046191	Nfa-8	Deletions (total of ~2100 bp) partially removing exons 1-4	gcttttgataagacccaaaataggc, tttagctagctgtcttcactatcg
At5g26617	5	9360279	9360673	Tamm-2	Deletion of 360 bp partially removing exon 1	actccttcaggcgattatacg, tttgtgaactctgaagagaagc
At5g26642	5	9272040	9275715	Ler-1	Deletion of 3839 bp removing entire gene	ataagaaatctcatcgacggc, gacgaagaagaagaggagacg
At5g28190	5	10167871	10171855	Fei-0	Deletion of 4155 bp removing exons 1-3, partially removing exon 4	acttgaaatgctttatcccg, gtacacaacacaagacataatctcc
At5g28210	5	10188569	10189090	Fei-0	Deletions of 168, 341 bp partially removing exon 1	aatacgtaaacctacgtgcgg, agagcattatcatcatcgtcg
At5g28646	5	10676726	10677360	L*er*-1	Deletion of 679 bp removing exons 5 and 6	tgtgttgaagaatcttggtgc, cttatgaaggcgagcataacc
At5g28823	5	10838025	10838824	Van-0	Deletion of 776 bp partially removing exons 2 and 3	ttcaaagcatgacaactaggg, ttacttagagcatggaacccc
At5g28930	5	10971033	10972860	Br-0	Deletion of 1818 bp removing exons 11-15	aatatccgggcattttaacc, ggtaaatttcagtaacgaaggg
At5g35230	5	13508666	13509037	Sha	Deletion of 209 bp partially removing exon 1	gcagttgaagcagttcttgg, atttggcggaaatgaagc
At5g36870	5	14538335	14538928	L*er*-1	Deletion of 563 bp removing exons 11-12, partially removing exon 10	ctgcaaaccttagattcatgc, cagattcactggtttcattcc

At5g37160	5	14724366	14725048	Br-0	Deletion of 671 bp partially removing exon 4	acttgtggatggaagaagagg, tgacggttactgagaaatccc
At5g37310	5	14790969	14791596	Tsu-1	Polymorphic region of 948 bp (PMN=95) overlapping exons 4 and 5	attggtgggatctctcttgg, ggtgatgtatttaggttcccc
At5g37760	5	15015394	15015899	Tamm-2	Deletion of 435 bp partially removing exon 4	ggagaagaaaaagctgattgg, cggttgtttctatatcttctgc
At5g38680	5	15496224	15496573	Rrs-7	Polymorphic region of 837 bp (PMN=105) overlapping exon 2	gtgtagtgctccaaaagatgg, gagtaattgatgctgcactgg
At5g38690	5	15496224	15496573	Rrs-7	Polymorphic region of 837 bp (PMN=105) overlapping exon 13	gtgtagtgctccaaaagatgg, gagtaattgatgctgcactgg
At5g39390	5	15781854	15783095	C24	Deletions (total of ~1247 bp) partially removing exons 1, 2, and 3	gggtcatgacaataaacatgc, ggaagagattttcaggttccc
At5g41950	5	16806566	16806962	Cvi-0	Deletion of 441 bp removing exon 14	gctggtctctgcattgatacc, tgaacttaggatacacgcacc
At5g41960	5	16806566	16806962	Cvi-0	Polymorphic region of 390 bp (PMN=75) overlapping exon 1	gctggtctctgcattgatacc, tgaacttaggatacacgcacc
At5g42965	5	17253923	17254359	Rrs-7	Deletion of 461 bp removing entire gene	cgcagaaactacatggaacc, tctcaatgacattctggatgg
At5g43550	5	17514856	17515162	Ts-1	Polymorphic region of 819 bp (PMN=140) overlapping exon 1	ggctacgagcaagtagactcc, cgcaacttagattcacaatagg
At5g43940	5	17703643	17703948	Lov-5	Polymorphic region of 711 bp (PMN=73) overlapping exon 9	agatatcaactcgtccgttcc, tgaagtatgagattgttgcgg
At5g43950	5	17703643	17703948	Lov-5	Polymorphic region of 711 bp (PMN=73) overlapping exon 2	agatatcaactcgtccgttcc, tgaagtatgagattgttgcgg
At5g44510	5	17946607	17947531	Nfa-8	Polymorphic region of 515 bp (PMN=75) overlapping exon 7	tttctcgatctggattgtagg, gttctgtacacaccaagcagc
At5g44850	5	18125258	18125655	Sha	Deletion of 533 bp partially removing exon 1	ctttcaacgacaagaacaagc, agtttctcaagaacagagcagc
At5g45050	5	18198359	18198698	Bay-0	Polymorphic region of 723 bp (PMN=43) overlapping exon 1 and 2	ctccagcaaacaataaacacc, gagcaaatcgtctacatcagc
At5g45095	5	18224708	18226287	Fei-0	Deletion of 1724 bp removing entire gene	cagtccgattcaatatgtaatgtcc, aagtgtttaaccacaacacgg
At5g45220	5	18316298	18316483	Bay-0	Polymorphic region of 288 bp (PMN=44) overlapping exon 6	gtacaagcttgaagagcatcc, tgatgagcctatgataaagcg
At5g46120	5	18719764	18720107	L*er*-1	Deletion of 335 bp partially removing exon 1	aaaatctcagatacgagtgacg, aggtccattaagatccactgc
At5g48770	5	19790008	19797277	Van-0	Deletions (total of ~7400 bp) removing majority of gene*	gagattgatatacgaacccgc, aagacacttcttccaaagatgg
At5g48780	5	19790008	19797277	Van-0	Deletions (total of ~7400 bp) removing majority of gene*	gagattgatatacgaacccgc, aagacacttcttccaaagatgg
At5g49020	5	19889847	19890835	Sha	Deletions (total of ~950 bp) removing exons 6, 8-9, partially removing 7 and 10*	tacaggattatgtgaggacgg, aattctggtacacacagacgc
At5g49290	5	20000089	20000531	Br-0	Polymorphic region of 915 bp (PMN=132) overlapping exon 6	gtctcacaacaaatttagcgg, agtagctgttctcgtcgaagg
At5g51195	5	20820665	20821849	Bur-0	Deletion of 1330 bp removing exon 1 and 2	accatccagacttgctagacg, gaagagagggaaatgagttgc
At5g53050	5	21528135	21528513	Lov-5	Polymorphic region of 794 bp (PMN=94) overlapping exons 8 and 9	ctatgaaatgtgcaggagagg, atgcagacatgatgtgattgc

Table S12. Multiple regression analysis results of the genomic features that best account for variability in nucleotide diversity across 50 kb windows.

	Estimate[a]	Std. Error	*t* value	*p*-value
Intergenic diversity				
Distance to centromere	-9.1×10^{-11}	7.5×10^{-12}	-12.1	$< 2\times10^{-16}$
GC content (all sites)	-2.5×10^{-2}	3.0×10^{-3}	-8.2	3.3×10^{-16}
Number of NB-LRR genes	4.2×10^{-4}	5.1×10^{-5}	8.2	4.8×10^{-16}
Amount of missing data (intergenic sites)	2.1×10^{-3}	2.8×10^{-4}	7.3	4.3×10^{-13}
Number of repetitive probes (all sites)	1.6×10^{-3}	2.2×10^{-4}	7.1	2.2×10^{-12}
GC content (intergenic sites)	8.2×10^{-3}	2.6×10^{-3}	3.1	2.0×10^{-3}
Four-fold degenerate diversity				
GC content (all sites)	-6.0×10^{-2}	4.0×10^{-3}	-14.9	$< 2\times10^{-16}$
Distance to centromere	-2.1×10^{-10}	2.1×10^{-11}	-9.8	$< 2\times10^{-16}$
Missing data (four-fold sites)	1.4×10^{-2}	1.7×10^{-3}	8.4	$< 2\times10^{-16}$
Repetitive probes (four-fold sites)	-1.3×10^{-2}	1.6×10^{-3}	-8.0	2.0×10^{-15}
Missing data (all sites)	8.9×10^{-3}	1.2×10^{-3}	7.3	4.4×10^{-13}
Repetitive probes (all sites)	-6.1×10^{-3}	1.1×10^{-3}	-5.7	1.5×10^{-8}
Number of NB-LRR genes	8.1×10^{-4}	1.4×10^{-4}	5.6	2.0×10^{-8}
GC content (four-fold sites)	1.1×10^{-2}	2.1×10^{-3}	5.1	3.5×10^{-7}
Number of all genes	1.1×10^{-4}	2.4×10^{-5}	4.4	1.0×10^{-5}

[a] GC content, repetitive probes, and missing data were all measured as a proportion of sites and so range from 0 to 1; number of NB-LRR genes or all genes are actual counts, and distance to the centromere was measured in base pairs (bp).

Table S13. Genes in chromosome 1 candidate sweep region.

Gene	Gene type	Description[a]
AT1G54450	Protein coding	Calcium-binding EF-hand family protein
AT1G54460	Protein coding	Expressed protein
AT1G54470	Protein coding	Encodes a Cf-like gene
AT1G54480	Protein coding	Encodes a Cf-like gene
AT1G54490	Protein coding	5'-3' exoribonuclease (XRN4), identical to XRN4
AT1G54500	Protein coding	Rubredoxin family protein
AT1G54510	Protein coding	Protein kinase family protein
AT1G54520	Protein coding	Expressed protein
AT1G54530	Protein coding	Calcium-binding EF hand family protein
AT1G54540	Protein coding	Expressed protein
AT1G54550	Protein coding	F-box family protein
AT1G54560	Protein coding	Myosin, putative
AT1G54570	Protein coding	Esterase/lipase/thioesterase family protein
AT1G54575	Protein coding	Expressed protein
AT1G54580	Protein coding	Acyl carrier protein, chloroplast, putative
AT1G54590	Protein coding	Splicing factor Prp18 family protein
AT1G54600	Pseudogene	Pseudogene
AT1G54610	Protein coding	Protein kinase family protein
AT1G54620	Protein coding	Invertase/pectin methylesterase inhibitor family protein
AT1G54630	Protein coding	Acyl carrier protein 3, chloroplast (ACP-3)
AT1G54640	Protein coding	F-box family protein-related
AT1G54650	Protein coding	Expressed protein
AT1G54660	pseudogene	Pseudogene, similar to vetispiradiene synthase
AT1G54670	Pre tRNA	tRNA-Ala (anticodon: TGC)
AT1G54680	Protein coding	Expressed protein
AT1G54690	Protein coding	Histone H2A, putative
AT1G54700	Protein coding	Hypothetical protein
AT1G54710	Protein coding	Expressed protein, contains 3 WD-40 repeats
AT1G54720	Protein coding	Early-responsive to dehydration protein-related / ERD protein-related, similar to ERD6 protein
AT1G54730	Protein coding	Sugar transporter, putative, similar to ERD6 protein
AT1G54740	Protein coding	Expressed protein
AT1G54750	Pseudogene	Pseudogene
AT1G54760	Protein coding	MADS-box family protein
AT1G54770	Protein coding	Expressed protein
AT1G54780	Protein coding	Thylakoid lumen 18.3 kDa protein
AT1G54790	Protein coding	GDSL-motif lipase/hydrolase family protein
AT1G54820	Protein coding	Protein kinase family protein
AT1G54830	Protein coding	CCAAT-box binding transcription factor Hap5a, putative
AT1G54840	Protein coding	Expressed protein
AT1G54850	Protein coding	Expressed protein

AT1G54860	Protein coding	Expressed protein
AT1G54870	Protein coding	Similar to short-chain dehydrogenase/reductase (SDR) family protein
AT1G54880	Protein coding	Hypothetical protein
AT1G54890	Protein coding	Late embryogenesis abundant protein-related / LEA protein-related
AT1G54905	Pseudogene	Copia-like retrotransposon family
AT1G54920	Protein coding	Expressed protein
AT1G54923	Protein coding	Expressed protein
AT1G54926	Protein coding	Expressed protein
AT1G54930	Protein coding	Zinc knuckle (CCHC-type) family protein
AT1G54940	Protein coding	Glycogenin glucosyltransferase (glycogenin)-related

[a] Modified from TAIR (*S12*).

Table S14. Genes in chromosome 5 candidate sweep region.

Gene	Gene type	Description[a]
AT5G08610	Protein coding	DEAD box RNA helicase (RH26), strong similarity to RNA helicase RH26
AT5G08620	Protein coding	DEAD box RNA helicase (RH25)
AT5G08630	Protein coding	DDT domain-containing protein
AT5G08640	Protein coding	Flavonol synthase 1 (FLS1)
AT5G08650	Protein coding	GTP-binding protein LepA, putative
AT5G08660	Protein coding	Expressed protein
AT5G08670	Protein coding	ATP synthase beta chain 1, mitochondrial
AT5G08680	Protein coding	ATP synthase beta chain, mitochondrial, putative
AT5G08690	Protein coding	ATP synthase beta chain 2, mitochondrial
AT5G08710	Protein coding	Regulator of chromosome condensation (RCC1) family protein / UVB-resistance protein-related
AT5G08712	miRNA	Encodes a microRNA. Targets At1g52150. Mature sequence: TCGGACCAGGCTTCATTCCCC
AT5G08717	miRNA	Encodes a microRNA. Targets At1g53160. Mature sequence: TCGGACCAGGCTTCATTCCCC
AT5G08720	Protein coding	Expressed protein
AT5G08730	Protein coding	IBR domain-containing protein
AT5G08740	Protein coding	Pyridine nucleotide-disulphide oxidoreductase family protein
AT5G08750	Protein coding	Zinc finger (C3HC4-type RING finger) family protein
AT5G08770	Protein coding	Expressed protein
AT5G08780	Protein coding	Histone H1/H5 family protein
AT5G08790	Protein coding	No apical meristem (NAM) family protein

[a] Modified from TAIR (*S12*).

Table S15. Field descriptions for Perlegen resequencing traces.

TRACE_NAME	Unique identifier for this trace, composed by concatenating the TEMPLATE_ID and TRACE_END.
TEMPLATE_ID	Uniquely identifies a pair of traces for forward and reverse tilings of the same sequence interval from the same scan: composed from the RUN_GROUP_ID, the scan date, and a code identifying the interval of tiled sequence.
TRACE_END	The orientation of the tiled fragment for this trace ("F" for forward or "R" for reverse).
SUBSPECIES_ID	The strain name for the DNA sample used in this experiment.
RUN_GROUP_ID	An identifier that groups together all traces from the same scanned image, corresponding to a single GeneChip DAT file, and a single analysis run.
PREP_GROUP_ID	Groups together all scans from a single hybridization experiment, i.e., a single physical array. For wafer-scale hybridizations, many scans are made to cover an entire wafer, and a wafer may be hybridized with several samples with different fluorophores.
CHIP_DESIGN_ID or FEATURE_ID_FILE_NAME	Identifies the chip design for the array covered by this RUN_GROUP_ID.
REFERENCE_ACCESSION	NCBI GenBank accession for the source sequence used for design of the array for this tiled interval
REFERENCE_OFFSET	Position in the GenBank sequence corresponding to the first tiled base in this trace file.