Detecting Polymorphic Regions in *Arabidopsis thaliana* with Resequencing Microarrays

G. Zeller,[1,2] R.M. Clark,[2,†] K. Schneeberger,[2,†] Anja Bohlen,[1] D. Weigel,[2] G. Rätsch[1,*]

[1] Friedrich Miescher Laboratory of the Max Planck Society, Tübingen, Germany

[2] Max Planck Institute for Developmental Biology, Department of Molecular Biology, Tübingen, Germany

March 6, 2008

† authors contributed equally
∗ corresponding author (Gunnar.Raetsch@tuebingen.mpg.de)

Supplemental Methods

Evaluation on representative genomic sequences

We aligned genomic sequences available for accessions L*er*-1, C24, and Cvi-0 to the Col-0 reference genome sequence to produce evaluation data sets for genome-wide PR predictions. For L*er*-1 we used shotgun sequence contigs from the Monsanto *A. thaliana* resequencing project [Jander et al., 2002] available at TAIR. Only contigs of length ≥ 1 kb and containing called nucleotides (i.e., A,C,G,T) were included in subsequent analyses. Using BLAT [Kent, 2002], with parameters `tileSize=10` and `minIdentity=80`, we aligned the Monsanto contigs to the Col-0 reference genome. Given the shotgun nature of the data (about 2-fold redundant; [The Arabidopsis Genome Initiative, 2000]), we applied several filters to remove potentially misassembled contigs and misalignments. First, we removed alignments which contained L*er*-1 deletions of length 100 nt or more. This was motivated by the observation that the alignments contained a high proportion of very large gaps most of which are likely due to assembly errors in the L*er*-1 contigs. The bias resulting from this filter on performance assessment is expected to be negligible as in the 2010 data 99.4% of all deletions are smaller than 100 nt. Relaxing these filter criteria to a maximal deletion length of 1,000 nt only marginally changed the sensitivity and specificity estimates (at most 1%). Finally, we also excluded Monsanto contigs for which more than one high identity match to the reference genome was observed; only if the second best BLAT match had at least 20% lower identity than the best match was considered, and only the best matches meeting this criterion were used for subsequent analyses.

For Cvi-0 and C24, we aligned finished BAC clone sequences (accession numbers EF637083 and EF182720, respectively; [Sherman-Broyles et al., 2007, Tang et al., 2007]) spanning the S-locus region to the reference genome sequence with the alignment program stretcher in the EMBOSS package [Rice et al., 2000]. Alignments were then manually corrected to give a total of 51 kb of aligned sequence from both clones.

From the resulting sets of genomic sequence alignments, we extracted SNPs and indels to construct label PRs, and we assessed sensitivity and specificity (see Methods). Among the L*er*-1, Cvi-0, and C24 data sets, specificity, which is not expected to be strongly affected by errors in the genomic sequence data, varied comparatively little (Supplemental Table S4 and Supplemental Table S3). However, sensitivity was markedly lower for the shotgun L*er*-1 data. To assess whether sequence errors in the L*er*-1 contigs/alignments were affecting the estimate of sensitivity, we compared PR labels in regions where the L*er*-1 genomic contigs overlapped 2010 sequence data for L*er*-1. In these overlapping regions, which consisted of 269 kb, we also compared PR predictions to PR labels from the 2010 set and to those extracted from the L*er*-1 genomic data (Supplemental Table S3). The large disagreement between the two sets of PR labels, as well as the discrepancy between sensitivity estimates for the different labels, indicated that a substantial proportion of apparent polymorphisms in the genomic data resulted from either sequencing or assembly errors in the shotgun Monsanto data. We therefore multiplied the sensitivity estimate for L*er*-1 predictions obtained from the genomic data by the resulting fold difference in sensitivity estimates for predictions evaluated on 2010 and on the genomic sequences for which the data sets overlapped (a factor of about 1.5; Supplemental Table S3). Both the uncorrected (u) and corrected (c) estimates for sensitivity for the genome-wide L*er*-1 predictions are given Supplemental Table S4.

Ability of mPPR to predict long deletions

We assessed the ability of our method to detect long deletions, which were absent in 2010, by using a test set of known deletions in the AtAD20 accessions [Clark et al., 2007]. We examined deletions >300 bp, which corresponded to 127 deletions of lengths between 302 and 10,536 bp (in total 118,566 deleted bases were examined across all 19 target accessions). Of the known deleted bases, 86.8% were included within PR boundaries in the appropriate accession (Supplemental Table S5). Where deleted bases were not included, 38.7% were repetitive as defined by *RM* (see above), a 2.1-fold over-representation relative to the genome average (Supplemental Table S5 and [Clark et al., 2007]). The deletions we employed for validation were initially identified using array methods, and likely represent a comparatively simple prediction task (e.g., comparatively low repeat content; see [Clark et al., 2007] for a discussion). Minimally, however, our method was highly effective at identifying the approximate locations of long deletions polymorphisms in unique sequences (see also Fig. 4and Supplemental Fig. SS6).

Experimental characterization of predictions

We used PCR and dideoxy sequencing to characterize predictions at the *RPM1* locus for which high polymorphism had been reported previously [Grant et al., 1998] [Shen et al., 2006]. Genomic DNA was prepared from three week old seedlings with standard methods. For PCR, primers flanking *RPM1* were design using Primer 3.0 [Rozen and Skaletsky, 2000]; the predictions themselves were used to select primer pairs likely to hybridize to target sequences without mismatches (see Supplemental Table S8 for primer sequences and designations). PCR reaction mixtures consisted of 20 mM Tris-HCL, 100 mM KCL, 1.5 μM MgCl$_2$, 200 μM of each dNTP, 0.5 unit Phusion polymerase, 1 μM of each primer, and $\sim$100 ng genomic DNA. Thermocycling was performed with a BIORAD DNA Engine Thermal Cycler (BIORAD, city, state) as follows: 98°C for 0:30 min, 30 cycles of 98°C for 0:08 min, 60.3°C for 0:30 min, and 72°C for either 1:00 or 2:30 min, and then 72°C for 8 min. The elongation step was for 1:00 min for primer pair 5 and for 2:30 min for all other primer pairs. The sizes of amplicons were established by electrophoresis on 1% agarose gels, and primary amplification products were purified from gel slices using the Promega WIZARD SV Gel Extraction Kit (Promega, city, state). Sequencing of purified products was performed with Big Dye Terminator chemistry (Applied Biosystems, Foster City, CA) using an ABI Prism 3730 capillary sequencer (Applied Biosystems). The software DNAstar Seqman was used for vector clipping. Sequence reads for each accession were aligned to the reference genome sequence using the program MUSCLE [Edgar, 2004] with a gap open penalty of 1000 and a gap extension penalty of 10^{-6}. Alignments were then refined manually.

Evaluation of genome-wide polymorphism levels

We assessed genome-wide patterns of polymorphism along each chromosome with sliding windows of size 100,001 bp (Fig. 5 and Supplemental Fig. SS7). Using the PR data, we calculated a measure of polymorphism defined by the fraction of positions in a window that were included within a PR in any accession. We calculated an analogous measure for the SNP data in MBML2 [Clark et al., 2007].

Polymorphism estimates for noncoding regions

We determined polymorphism for the 1000 bp upstream to the transcription start and downstream to transcription termination sites for coding genes based on the TAIR6 genome annotation. Polymorphism at and nearby genes was calculated as the average percentage of accessions (excluding Col-0) harboring a PR prediction at the position. We then averaged the results across all genes, thereby standardizing on the transcription start and termination sites. For comparison, we calculated the analogous measure with MBML2 data. In the analysis, we only considered genes with annotated 5' and 3' UTRs. An analogous calculation was also applied to assess polymorphism levels around splice sites from positions-50 to $+50$ relative to the dinucleotide donors and acceptors. For this, we only considered genes with a single annotated splice form.

Relation of PRs to predicted *cis*-elements

Position-wise *cis*-element density was calculated using the predictions of [O'Connor et al., 2005] that were based on putative binding sites for 105 transcription factors (TFs). With permutation tests we assessed whether the overlap of PRs to *cis*-elements differed from that expected by chance for each accession. For this analysis, we considered the same set of genes as for the polymorphism analysis excluding genes with upstream regions shorter than 1000 bp. The tests were performed as follows. For each gene g in the set G of genes let C_g denote the set of positions of annotated *cis*-elements in the 1 kb promoter region of g and similarly P_g the set of positions in PR predictions in the same promoter. Then we defined the fraction of polymorphic regions overlapping with *cis*-elements in this promoter as $f_g = |C_g \cap P_g|/|C_g|$ and obtained the average over all genes in G as $\hat{f} = \frac{1}{|G|}\sum_{g\in G} f_g$.

We randomly permuted the association between *cis*-elements and PR predictions ($n = 1000$ times) to obtain $c_{g,h}^{(i)} = |C_g \cap P_h|/|C_g|$ where $h = perm_i(g)$ $\forall i = 1,\ldots n$ and calculated the average $\hat{c}^{(i)} = \frac{1}{|G|}\sum_{g\in G} c_{g,h}$. Figure 6 C shows an example of the histogram of 1000 such $\hat{c}$ values of one permutation test for accession Bor-4 (the arrow indicates where the original $\hat{f}$ value falls within the distribution of $\hat{c}^{(i)}$). We also calculated $rank(\hat{f}) = |\{i \mid \hat{c}^{(i)} < \hat{f}\}|$ which was found to be 0 in all 19 permutation tests (see also Supplemental Fig. SS8).

Annotation of Predictions Relative to Genes

We calculated the overlap of PRs to coding sequences based on the TAIR6 annotation with gene family descriptions as previously reported [Clark et al., 2007]. We performed a similar analysis for these genes on the basis of orthology to poplar. Orthology was established by using *inparanoid* (version 2.0) [Remm et al., 2001]. Each member of a group of orthologs was assessed to be orthologous to all genes of the other species of the same group.

When mapping PR predictions to miRNA genes, we used the following divisions: precursor end (miRNA arm), miRNA, loop region, miRNA*, precursor end (miRNA* arm) (see Fig. 7 C). Since the location of the miRNA* is not annotated in RFam [Griffiths-Jones et al., 2006], we calculated a secondary structure for each miRNA using RNAfold [Hofacker et al., 1994]. The star region was defined as the region binding to the annotated micro, shifted by two nucleotides to the 3' end of the miRNA. To account for length differences between miRNA genes, all were mapped to a prototypical miRNA gene consisting of the five sections of length l_r ($r \in \{1, \ldots, 5\}$). For each section we set l_r to half the (rounded) average section length across all miRNAs. When mapping the PR predictions to this prototype, positions in a section of length m_r in a given miRNA were rescaled by a factor $\alpha = l_r/m_r$.

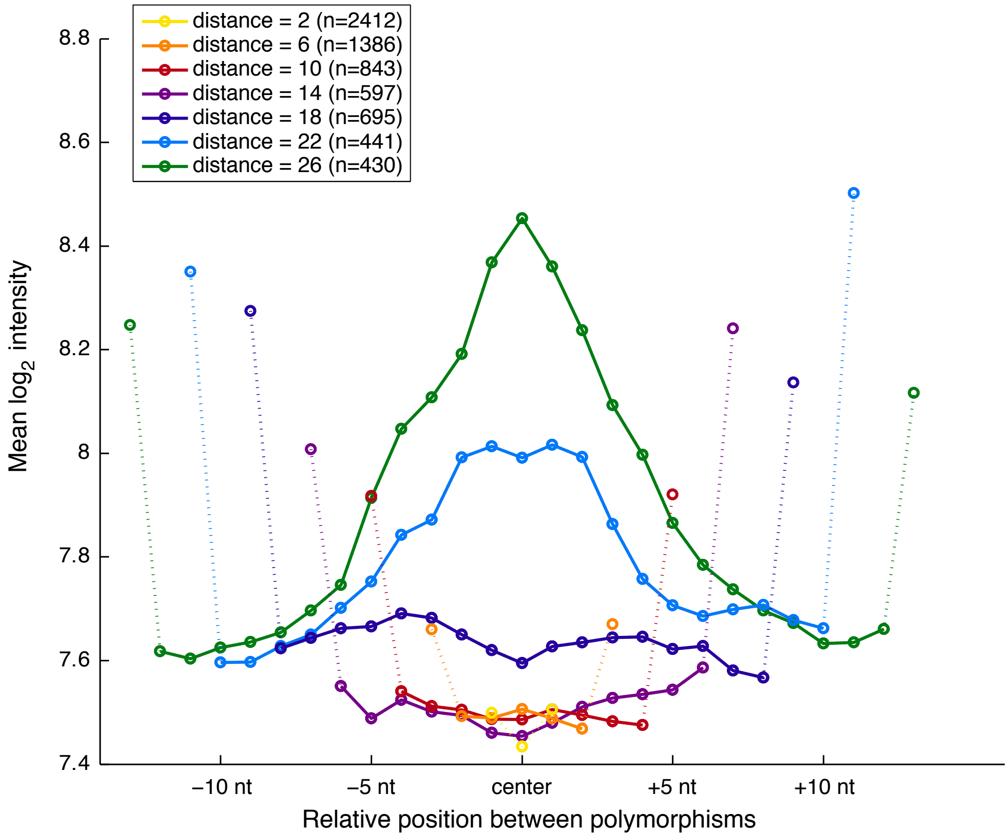

Figure S1: Intensities for features located between adjacent polymorphisms is reduced. Regions between polymorphisms at a distance ≤26 bp to each other were extracted and categorized according to this distance (see inset). For each distance category the maximal intensities for each probe quartet between polymorphisms were averaged for the forward and reverse strands resulting in a single curve per category (circles and solid lines). The outermost circles and dotted lines indicate the average intensities at polymorphic sites. All curves are centered and positions on the x-axis are relative to the center. Intensity at sites between polymorphisms ≤18bp from each other was generally suppressed. Intensities recovered for features between polymorphisms at greater distances (light blue and green curves). These findings motivated our use of 18 bp for defining PR and clustered SNPs (see main text).

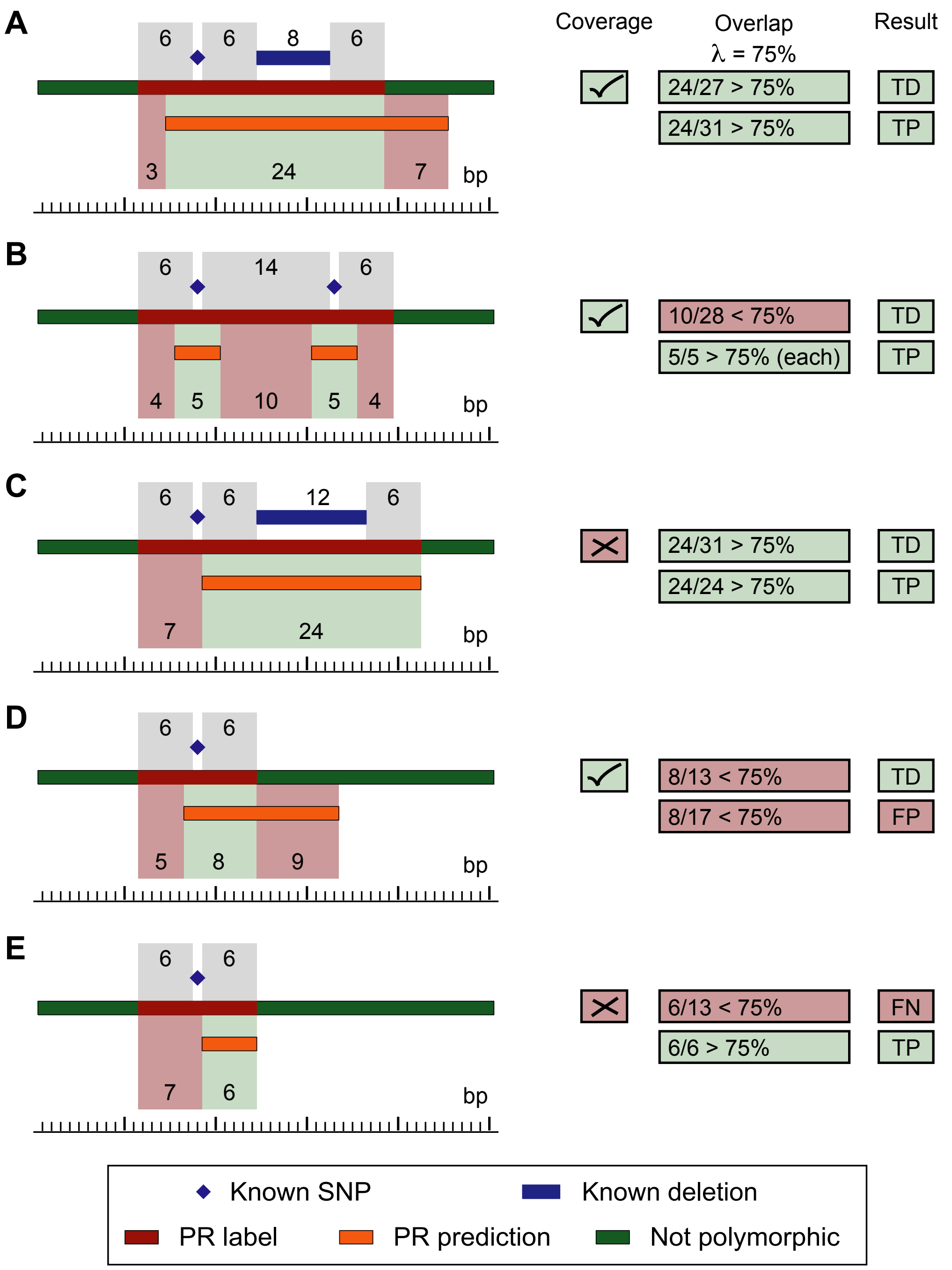

Figure S2: Illustration of performance assessment. To calculate sensitivity, i.e., for whether each label PR was a true discovery (TD, green shading) or a false negative (FN, red shading), we first checked if all underlying polymorphisms were included in one or more PR predictions (boxes with title "coverage"). If so, as for examples **A**, **B** and **D**, the label PR was counted as a TD. Otherwise, depending on whether a portion $\geq \lambda$ was overlapping with one or more PR predictions (boxes with title "overlap"), it was still counted as a TD (as in **C**), else as a FN (as in **E**). Specificity assessment was only based on the proportion of a PR prediction overlapping with label PRs. If a fraction $\geq \lambda$ of the prediction was also labeled as a PR, the prediction was counted as a true positive (TP, green shading, as in **A**, **B**, **C** and **E**), and otherwise as a false positive (FP, red shading, as in **D**).

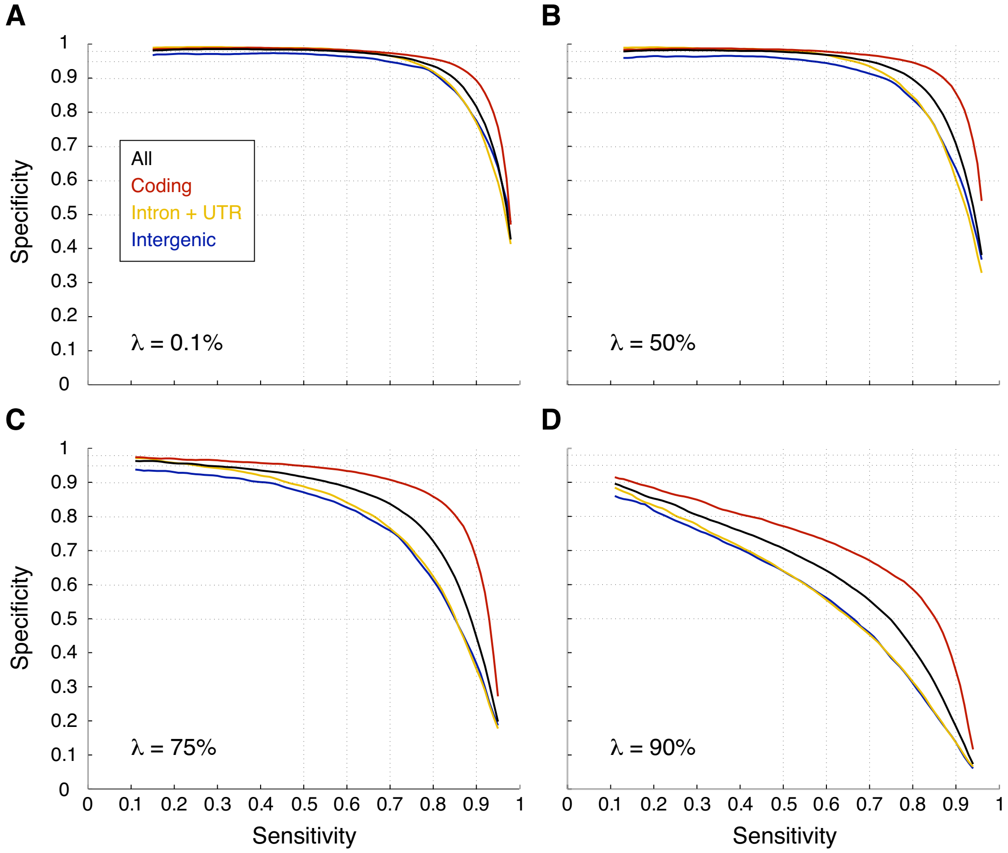

Figure S3: Dependency of performance on the choice of the minimal required overlap λ between known PRs and PR predictions. Shown are the sensitivity-specificity curves for 4 different choices of λ (panels **A**-**D**).

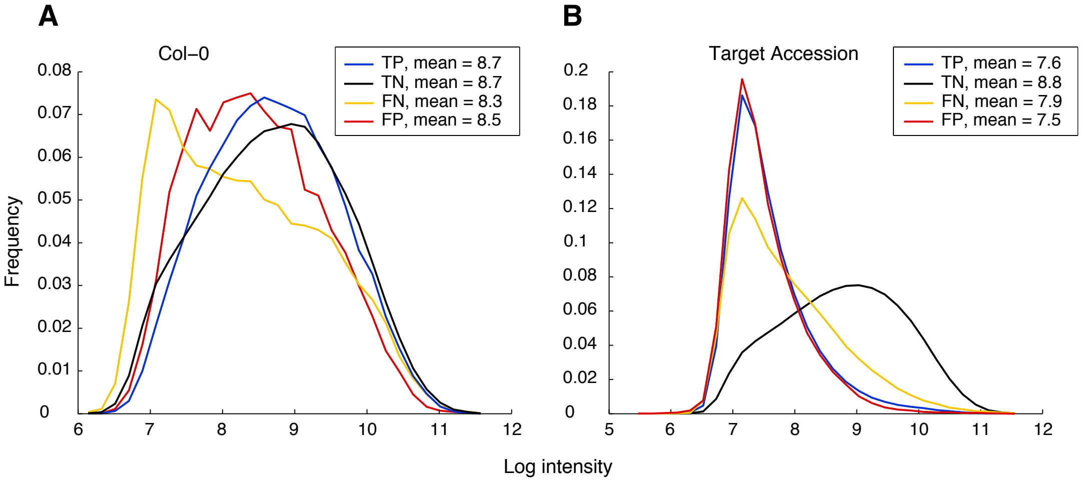

Figure S4: Detection of PRs is difficult in regions where hybridization intensities for the reference accession, Col-0, were reduced due to unfavourable hybridization properties of oligonucleotides. Intensity histograms were calculated separately for true positive (TP), false positive (FP), true negative (TN) and false negative (FN) sites from the maximum intensity in each probe quartet and were divided by the total counts to obtain frequencies on the ordinate (note different scales). **A**. Histograms for the reference accession, Col-0. **B**. Histograms for the accession in which PRs were predicted.

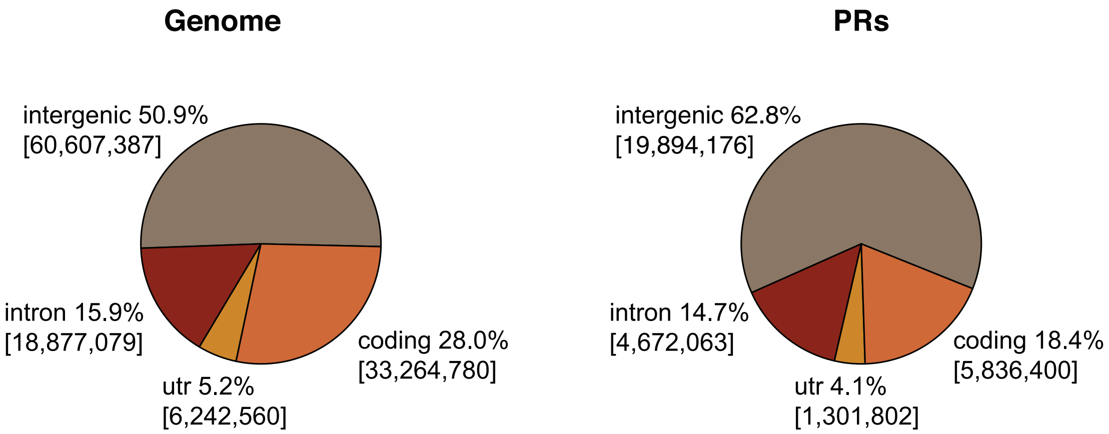

Figure S5: **A.** Arrayed bases by sequence type. **B.** Nonredundant bases included in PRs by sequence type.

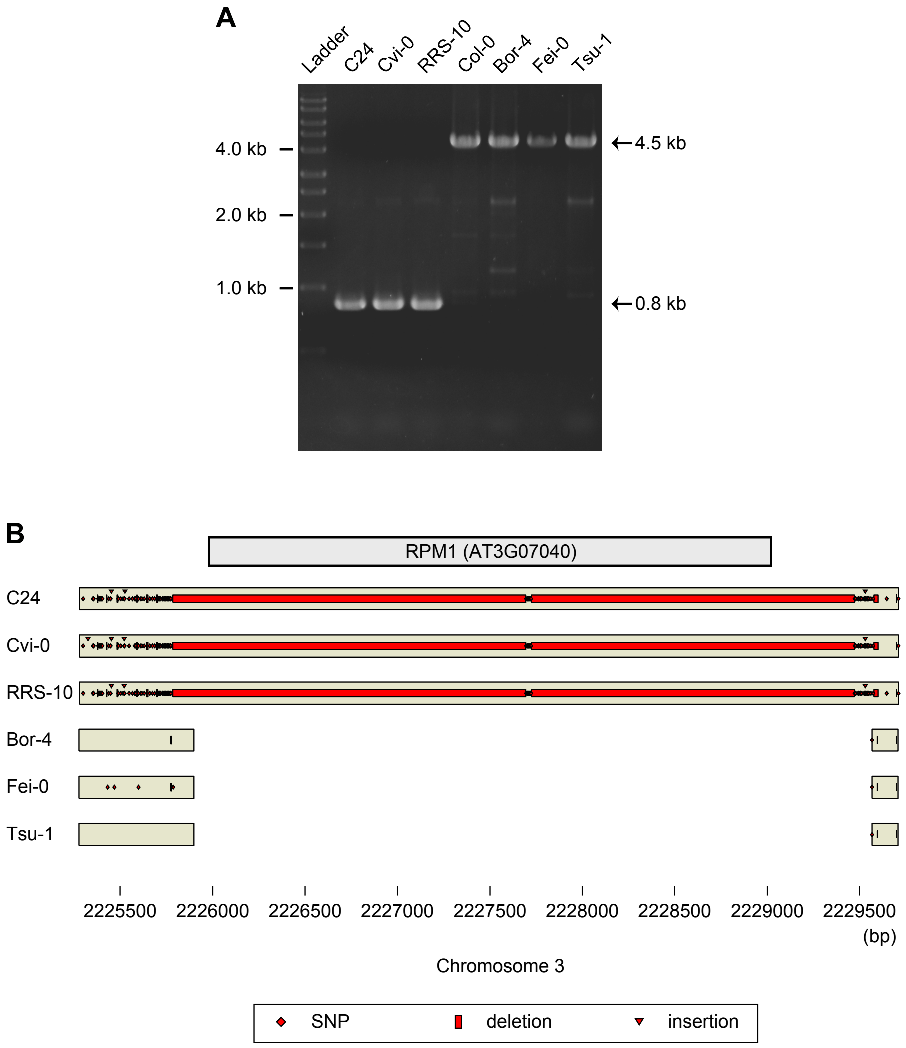

Figure S6: Underlying polymorphism at locations of PR predictions at the *RPM1* locus. **A**. Image of an ethidium bromide stained 1.0 % agarose gel showing PCR amplification products for seven accessions using primers flanking *RPM1* (Lane 1: DNA ladder; Lanes 2-8: PCR products for accessions as indicated at top). Products for three accessions (Bor-4, Fei-0, and Tsu-1; right) were of similar size to that of the Col-0 reference (center). For accessions C24, Cvi-0, and RRS-10, smaller products were observed. **B**. Schematic of polymorphisms inferred from end sequencing of primary amplification products shown in panel A (GenBank accession nos. ET181618 to ET181629). Chromosome and position is based on the reference sequence, and tan-colored boxes indicate where sequence data was obtained for each accession. For the smaller PCR products (C24, Cvi-0, and RRS-10; see panel A), complete sequence was obtained across amplicons, revealing many sequence changes compared to other accessions (polymorphism types are indicated at bottom). PR predictions for C24, Cvi-0, and RRS-10 (see Fig. 4) corresponded to large deletions at *RPM1* or to dense clusters of SNPs and small indels flanking the transcribed *RPM1* sequence. A small number of polymorphisms were also identified for Bor-4, Fei-0, and Tsu-1, many of which were also captured by PR predictions (see Fig. 4).

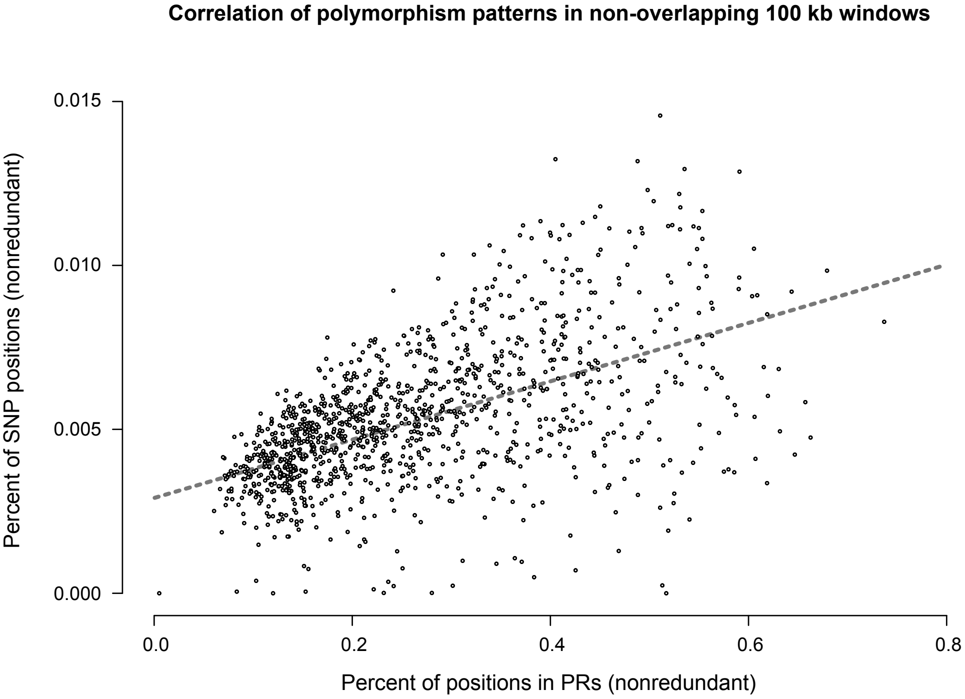

Figure S7: Correlation between estimates of polymorphism from PR predictions and MBML2 SNPs. Polymorphism was calculated as in Figure 5 for positions central to non-overlapping 100 kb windows, and estimates from the two data sets are significantly correlated (Pearson's cor = 0.54, P-value $< 10^{-15}$), even though the estimates sometimes differ substantially (see also Fig. 5). In these cases, polymorphism estimated from the PR data is often disproportionately higher. This finding is generally consistent with known ascertainment biases in the data sets. Regions of very high polymorphism are well delimited in the PR data, but are too divergent for explicit SNP prediction (i.e., they would largely be absent from MBML2; see also Table 2). Furthermore, PR predictions capture indel polymorphisms, including long deletions, and such predictions would lead to elevated estimates of polymorphism in the PR data relative to the SNP data. Ascertainment biases in both data sets, however, likely also contribute to differences in polymorphism estimates (e.g., for repetitive regions).

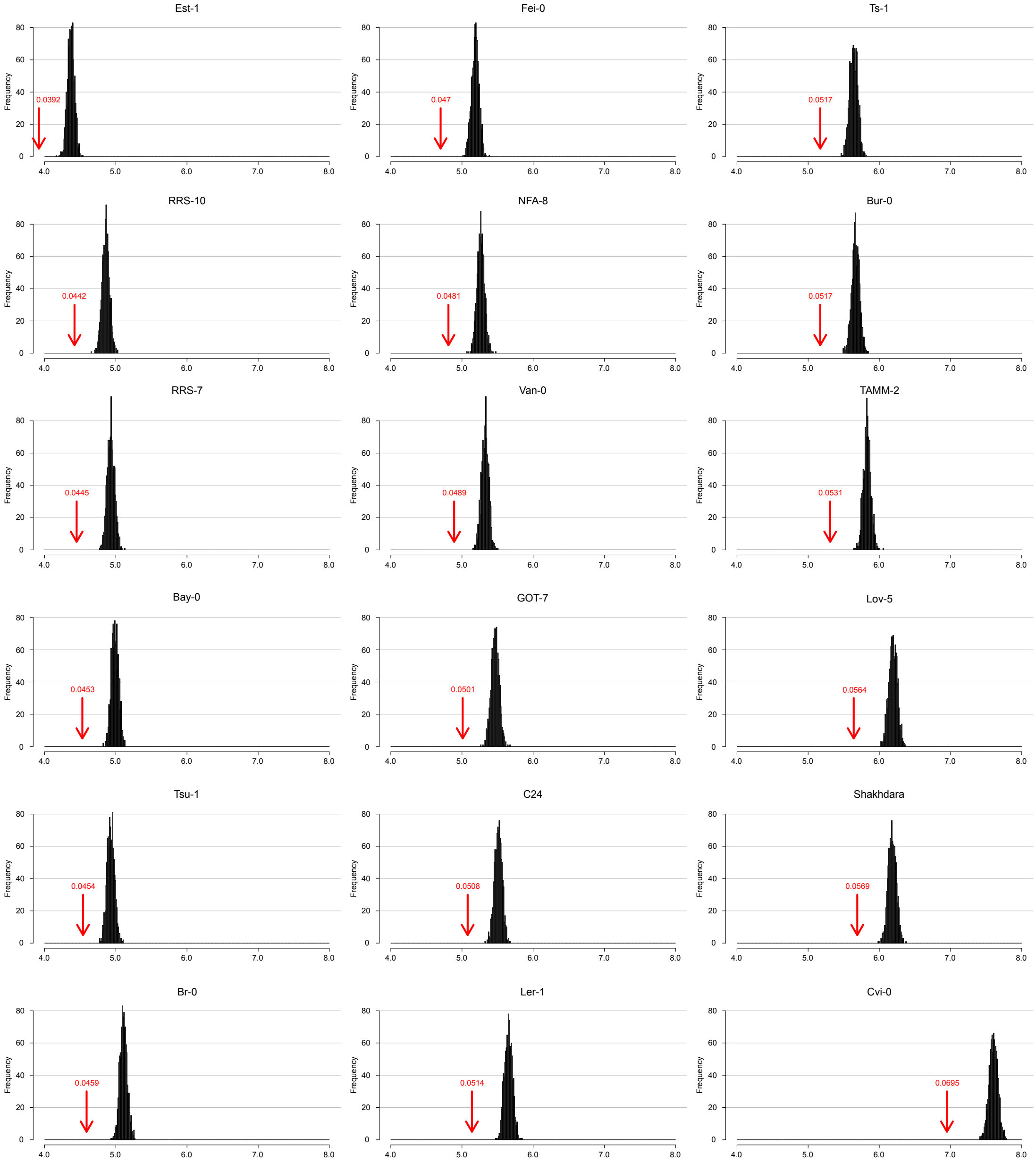

Figure S8: Percent overlap of PRs to *cis*-element motifs mapped to the *A. thaliana* genome for 18 accessions. See Figure 6 C legend and Supplemental Methods for details, and Figure 6 C for the analogous data for accession Bor-4. The panels are ranked top to bottom and left to right by the observed percent overlap.

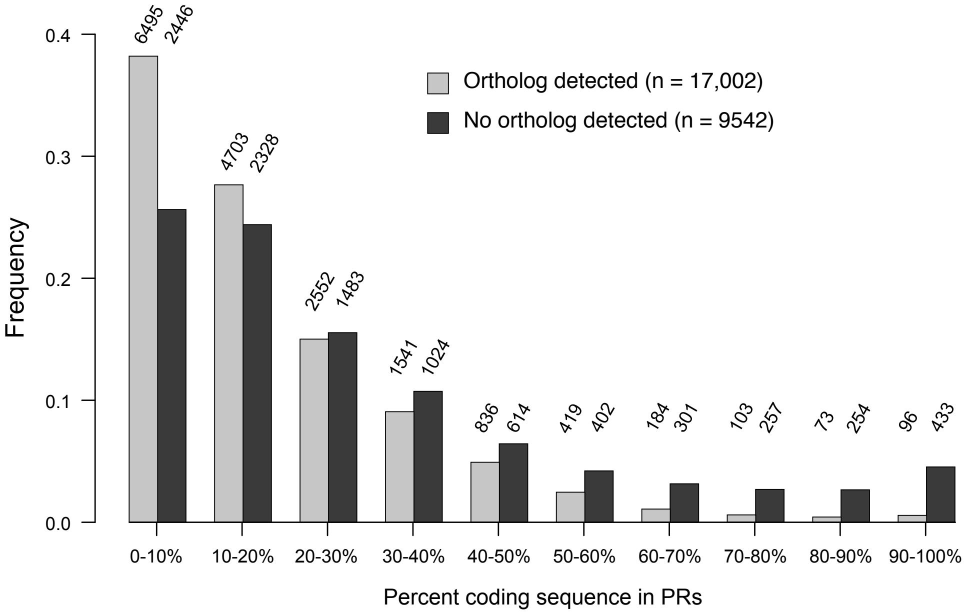

Figure S9: Percent nonredundant PR overlap to coding genes by orthology to black cottonwood. The TAIR6 *A. thaliana* annotation was used for the comparison, and is the basis for listed sample sizes (n). Orthology was established as described in the Supplemental Methods.

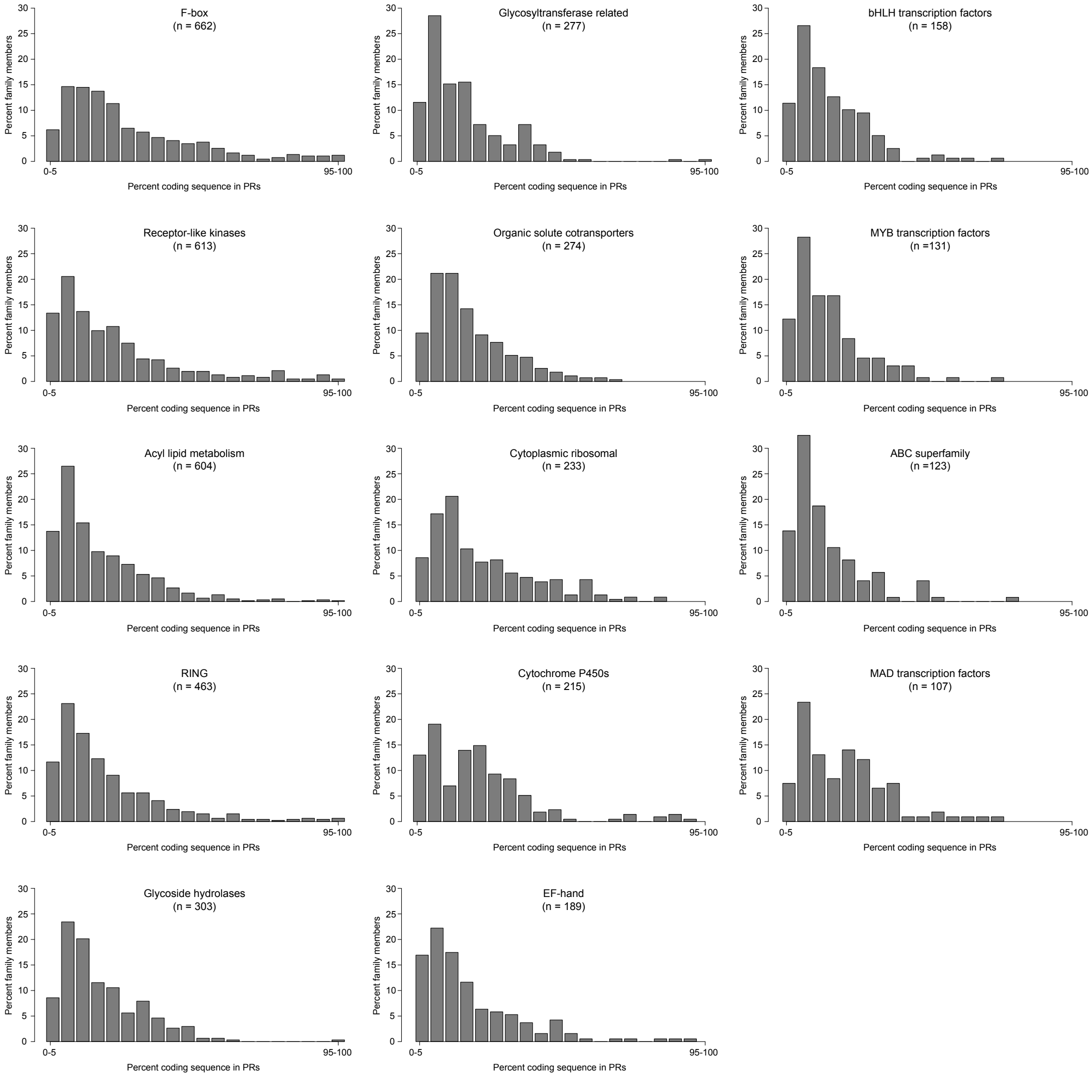

Figure S10: Distribution of coding genes by percent inclusion in PRs by gene family classification. See Figure 7 A-B for additional information and gene families.

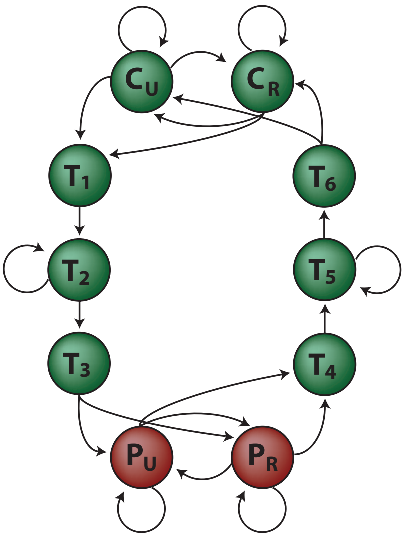

Figure S11: State-transition model. States are drawn as colored circles, transitions as arrows. $\mathtt{C}_\mathtt{U}$ and $\mathtt{C}_\mathtt{R}$ model conserved sites (unique and repetitive, respectively). Similarly, $\mathtt{P}_\mathtt{U}$ and $\mathtt{P}_\mathtt{R}$ model sites that are polymorphic or nearby a polymorphism. Additional states $\mathtt{T}_i$ model the *gradual* change in hybridization signal between conserved and polymorphic regions. The color of each state indicates the corresponding label (cf. Fig. 1).

Accession	Number of PRs	% genome in PRs	Specificity	Sensitivity
Bay-0	271,644	6.3	92.2%	54.9%
Bor-4	276,256	6.1	91.7%	55.6%
Br-0	276,913	6.5	88.4%	53.8%
Bur-0	284,143	6.6	93.0%	52.2%
C24	293,558	6.7	93.7%	52.7%
Cvi-0	361,184	8.5	87.1%	57.3%
Est-1	240,538	5.3	92.0%	49.9%
Fei-0	277,788	6.4	88.1%	55.4%
GOT-7	284,596	6.5	85.9%	55.9%
L*er*-1	302,450	7.0	90.0%	59.0%
Lov-5	320,648	7.3	87.6%	60.8%
NFA-8	283,544	6.5	92.3%	56.2%
RRS-10	260,721	5.9	93.8%	55.7%
RRS-7	275,700	6.3	89.6%	55.6%
Shakhdara	304,471	7.4	90.6%	55.8%
TAMM-2	307,564	7.2	88.7%	54.2%
Ts-1	303,340	7.0	91.2%	57.4%
Tsu-1	272,438	6.2	92.9%	56.8%
Van-0	281,600	6.6	NA	NA

Table 1: Whole-genome PR predictions and performance by accession. Predictions are for 90% specificity on 2010 as assessed across all accessions excluding Van-0 (cf. Table 1, $\lambda = 75\%$). Specificity and sensitivity for each accession as determined from 2010 is also given. 2010 data for Van-0 was not available; nevertheless, we used HMSVMs trained across data from all other accessions to predict PRs in Van-0. The absence of test data precluded evaluation of specificity and sensitivity for the Van-0 accession (NA is "not applicable").

	Bases (kb)	PRs	PRPs	Single-SNP	Multi-SNP	Deletion	Insertion	Complex	Empty
2010	10,967	34,054	20,073	12,435 [62%]	4,584 [23%]	438 [2%]	242 [1%]	1,607 [8%]	767 [4%]
C24	14	125	65	27 [42%]	23 [35%]	4 [6%]	0 [0%]	9 [14%]	2 [3%]
Cvi-0	37	265	169	76 [45%]	60 [35%]	6 [4%]	2 [1%]	23 [14%]	2 [1%]
L*er*-1	37,871	203,611	76,020	41,052 [54%]	18,331 [24%]	2,257 [3%]	1,319 [2%]	9,771 [13%]	3,290 [4%]

Table 2: Polymorphisms in predicted PRs. We distinguished between PR predictions (PRPs) containing only a single SNP (Single-SNP), multiple SNPs (Multi-SNP), one or more deletions (Deletion), one or more insertion sites (Insertion), SNPs and indels in combination (Complex), or no known polymorphism at all (Empty).

	λ= 75%		λ= 50%	
	Specificity	Sensitivity	Specificity	Sensitivity
PRPs vs. 2010 PRs	90%	72%	97%	79%
PRPs vs. Monsanto PRs	90%	50%	97%	53%
2010 PRs vs. Monsanto PRs	97%	48%	99%	51%

Table 3: Performance evaluation of PRPs on regions overlapping between 2010 and Monsanto L*er*-1 sequences/contigs. The first two rows show performance assessments of PR predictions (PRPs) against PRs extracted from alignments of 2010 L*er*-1 sequences and against PRs extracted from aligned Monsanto contigs, respectively. The third row shows overlap comparisons between the two sets of PR labels.

	Bases (kb)	PRs	PRPs	$\lambda = 75\%$		$\lambda = 50\%$	
				Spec.	Sens.	Spec.	Sens.
C24	14	124	65	95%	40%	100%	45%
Cvi-0	37	259	169	87%	61%	96%	67%
L*er*-1 (u)	37,871	186,916	74,354	88%	32%	96%	34%
L*er*-1 (c)					48%		53%

Table 4: Evaluation on genomic sequences. Bases denotes the number of aligned bases, PRs the polymorphic regions extracted from these alignments and PRPs the predicted polymorphic regions for the corresponding regions. Specificity and sensitivity are given for two different overlap cut-offs (see Methods). L*er*-1 (u) shows specificity and sensitivity values from a direct comparison to the alignments of the Monsanto contigs to the Col-0 reference sequence. We corrected the sensitivity by a factor estimated from the discrepancies of the sensitivity rates in regions where the 2010 L*er*-1 sequences overlap to the Monsanto contigs [L*er*-1 (u)] (see Supplemental Methods and Supplemental Table S3).

	Non-repetitive	Repetitive
Known deleted bases (total)	109,118	9,448
Known deleted bases in PRs	99,527	3,400
Known deleted bases not in PRs	9,591	6,048

Table 5: Known deleted bases in 127 long deletions (>300 bp) included within PR prediction boundaries by repeat content.

1	$IM_t(p) = \frac{1}{2}\,[\,\log(I^+_{max}(p)) + \log(I^-_{max}(p))\,]$
2	$IR(p) = IM_t(p) - IM_{Col}(p)$
3	$IW(p) = \frac{1}{9} \sum_{\delta=-4}^{4} IR(p+\delta)$
4	$IN(p) = \frac{1}{2} \sum_{\delta \in \{-1,+1\}} (IM_t(p) - IM_t(p+\delta))$
5	$QN(p) = \frac{1}{4} \sum_{\delta \in \{-1,+1\}} \sum_{s \in \{+,-\}} (Q^s_t(p)/(1+Q^s_t(p+\delta)))$
6	$MM(p) = \sum_{\delta=-4}^{4} (mism_t(p+\delta) - mism_{Col}(p+\delta))$
7	$WL(p) = 1 + \log_2(wl(p))$
8	$RM(p) = [[p \in \mathcal{R}]]$

Table 6: Features used for polymorphic region prediction (maybe move to supplement). Here we use the notation from [Clark et al., 2007]: In general, superscripts + or $-$ denote the strand, subscripts *acc* the accession (where t is the target and Col the reference accession), and [[.]] the indicator function. $I^s_{max}(p)$ denotes the maximum intensity in the probe quartet which queries site p and strand s, $Q^s(p)$ the quality score assigned to that probe quartet, $mism_{acc}(p) = [[B^+_{acc}(p) = seq(p)]] + [[B^-_{acc}(p) = seq(p)]]$ a count of mismatches between raw base calls B and reference sequence seq at site p, and $\mathcal{R}$ the set of repetitive sites. Word length $wl(p)$ equals the number of *consecutive* sites around p where $B^s(p') = seq(p') \quad \forall s \in \{+,-\}$. (For further details see supplement of [Clark et al., 2007].) All features were standardized prior to training (mean and standard deviation were estimated on the training set).

State	Label				
	non-polymorphic	SNP	insertion	deletion	tolerance
$\mathtt{C}_{\mathtt{U}}, \mathtt{C}_{\mathtt{R}}$	0	0.5	0.5	1	0
$\mathtt{P}_{\mathtt{U}}, \mathtt{P}_{\mathtt{R}}$	$0.5 + 0.1d$	0	0	0	0
$\mathtt{T}_1, \ldots, \mathtt{T}_6$	$0.1 + 0.1d$	0.1	0.1	0.1	0

Table 7: Position-wise loss $\ell(p)$. We use a "tolerance" region, comprising non-polymorphic nucleotides in labeled blocks (up to 9bp upstream and downstream of polymorphisms), where neither `C` nor `P` states incur any loss. For non-polymorphic sites outside the tolerance region the loss also depends on the distance to the nearest polymorphism; this distance contribution is denoted by d. Let $dist(p)$ be the distance from position p to the nearest polymorphism. Then, $d(p) = 0$, if $dist(p) \leq 9$, else $d(p) = dist(p) - 9$, if $9 < dist(p) \leq 21$, and $d(p) = 12$, otherwise.

Primer name	Sequence
pair_1_8-9_AT3G07040_RPM1-left	TGGTTTCGGTTTAGCGACTC
pair_1_8-9_AT3G07040_RPM1-right	AAAGCAGGAGCTGATGAGGA

Table 8: Primers used for PCR amplification and sequencing (see main text)

References

[Clark et al., 2007] Clark, R., Schweikert, G., Toomajian, C., Ossowski, S., Zeller, G., Shinn, P., Warthmann, N., Hu, T., Fu, G., and Hinds et al., D., *et al.*, 2007. Common sequence polymorphisms shaping genetic diversity in Arabidopsis thaliana. *Science*, **317**(5836):338–342.

[Edgar, 2004] Edgar, R., 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**(1):113.

[Grant et al., 1998] Grant, M., McDowell, J., Sharpe, A., de Torres Zabala, M., Lydiate, D., and Dangl, J., 1998. Independent deletions of a pathogen-resistance gene in Brassica and Arabidopsis. *PNAS*, **95**(26):15843–15848.

[Griffiths-Jones et al., 2006] Griffiths-Jones, S., Grocock, R., van Dongen, S., Bateman, A., and Enright, A., 2006. miRBase: microRNA sequences, targets and gene nomenclature. *Nucl. Acids Res.*, **34**(suppl_1):D140–144.

[Hofacker et al., 1994] Hofacker, I., Fontana, W., Stadler, P., Bonhoeffer, L., Tacker, M., and Schuster, P., 1994. Fast folding and comparison of RNA secondary structures. *Monatshefte f. Chemie*, **125**(2):167–188.

[Jander et al., 2002] Jander, G., Norris, S., Rounsley, S., Bush, D., Levin, I., and Last, R., 2002. Arabidopsis map-based cloning in the post-genome era. *Plant Physiol.*, **129**(2):440–450.

[Kent, 2002] Kent, W., 2002. BLAT–the BLAST-like alignment tool. *Genome Res.*, **12**(4):656–64.

[O'Connor et al., 2005] O'Connor, T., Dyreson, C., and Wyrick, J., 2005. Athena: a resource for rapid visualization and systematic analysis of Arabidopsis promoter sequences. *Bioinformatics*, **21**(24):4411–4413.

[Remm et al., 2001] Remm, M., Storm, C., and Sonnhammer, E., 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons;. *J. Mol. Biol.*, **314**(5):1041–1052.

[Rice et al., 2000] Rice, P., Longden, I., and Bleasby, A., 2000. Emboss: The european molecular biology open software suite;. *Trends in Genetics*, **16**(6):276–277.

[Rozen and Skaletsky, 2000] Rozen, S. and Skaletsky, H., 2000. Primer3 on the WWW for general users and for biologist programmers. In S, S. K. and Misener, S., editors, *Bioinformatics Methods and Protocols: Methods in Molecular Biology*, pages 365–386. Humana Press, Totowa, NJ.

[Shen et al., 2006] Shen, J., Araki, H., Chen, L., Chen, J., and Tian, D., 2006. Unique evolutionary mechanism in r-genes under the presence/absence polymorphism in Arabidopsis thaliana. *Genetics*, **172**(2):1243–1250.

[Sherman-Broyles et al., 2007] Sherman-Broyles, S., Boggs, N., Farkas, A., Liu, P., Vrebalov, J., Nasrallah, M. E., and Nasrallah, J. B., 2007. S Locus Genes and the Evolution of Self-Fertility in Arabidopsis thaliana. *Plant Cell*, **19**(1):94–106.

[Tang et al., 2007] Tang, C., Toomajian, C., Sherman-Broyles, S., Plagnol, V., Guo, Y.-L., Hu, T. T., Clark, R. M., Nasrallah, J. B., Weigel, D., and Nordborg, M., *et al.*, 2007. The Evolution of Selfing in Arabidopsis thaliana. *Science*, **317**(5841):1070–1072.

[The Arabidopsis Genome Initiative, 2000] The Arabidopsis Genome Initiative, 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, **408**(6814):796–815.