Detecting polymorphic regions in *Arabidopsis thaliana* with resequencing microarrays

Georg Zeller,[1,2] Richard M. Clark,[2,3] Korbinian Schneeberger,[2,3] Anja Bohlen,[1] Detlef Weigel,[2] and Gunnar Rätsch[1,4]

[1]*Friedrich Miescher Laboratory of the Max Planck Society, Tübingen 72070, Germany;* [2]*Max Planck Institute for Developmental Biology, Department of Molecular Biology, Tübingen 72070, Germany*

Whole-genome oligonucleotide resequencing arrays have allowed the comprehensive discovery of single nucleotide polymorphisms (SNPs) in eukaryotic genomes of moderate to large size. With this technology, the detection rate for isolated SNPs is typically high. However, it is greatly reduced when other polymorphisms are located near a SNP as multiple mismatches inhibit hybridization to arrayed oligonucleotides. Contiguous tracts of suppressed hybridization therefore typify polymorphic regions (PRs) such as clusters of SNPs or deletions. We developed a machine learning method, designated margin-based prediction of polymorphic regions (mPPR), to predict PRs from resequencing array data. Conceptually similar to hidden Markov models, the method is trained with discriminative learning techniques related to support vector machines, and accurately identifies even very short polymorphic tracts (<10 bp). We applied this method to resequencing array data previously generated for the euchromatic genomes of 20 strains (accessions) of the best-characterized plant, *Arabidopsis thaliana*. Nonredundantly, 27% of the genome was included within the boundaries of PRs predicted at high specificity (≈97%). The resulting data set provides a fine-scale view of polymorphic sequences in *A. thaliana*; patterns of polymorphism not apparent in SNP data were readily detected, especially for noncoding regions. Our predictions provide a valuable resource for evolutionary genetic and functional studies in *A. thaliana*, and our method is applicable to similar data sets in other species. More broadly, our computational approach can be applied to other segmentation tasks related to the analysis of genomic variation.

[Supplemental material is available online at www.genome.org.]

Describing the complement of sequence variation within a species is a first step in linking genetic variation to phenotypes (The International HapMap Consortium 2005), and the development of methods for whole-genome polymorphism discovery has been a top priority in the life sciences (Shendure et al. 2004). Toward this goal, the generation of high-density, oligonucleotide microarrays suitable for whole-genome variation detection was a major technological breakthrough (e.g., Chee et al. 1996; Patil et al. 2001; Hinds et al. 2005). Such microarrays, hereafter referred to as resequencing arrays, employ a 1-bp tiling path to query bases relative to a known reference sequence. Each base is interrogated with eight features that consist of forward and reverse strand 25-mer oligonucleotide quartets. Within a quartet, oligonucleotides are identical to the reference sequence except at the central position, where each sequence possibility is represented. When hybridized to labeled genomic DNA, the highest signal intensity is expected for the perfect match oligonucleotide, thereby predicting the base in the corresponding target DNA sample. Large-scale polymorphism discovery using resequencing arrays was first performed in humans, identifying a large fraction of common single nucleotide polymorphisms (SNPs) in the global population (Patil et al. 2001; Hinds et al. 2005).

Although conceptually simple, detection of polymorphisms from resequencing array data is nonetheless a computational challenge (Cutler et al. 2001; Patil et al. 2001; Clark et al. 2007). For SNPs, relative differences in feature intensities at a polymorphic position indicate the base call, and hybridization is reduced for flanking features as a consequence of off-center mismatches (cf. Fig. 1A,B). The resulting hybridization pattern provides a "SNP signature" that has been exploited by several algorithms to predict SNPs from resequencing array data (Patil et al. 2001; Hinds et al. 2005; Clark et al. 2007). However, where multiple SNPs or insertion/deletion (indel) polymorphisms are closely adjacent (occur within the same 25-mer), all oligonucleotides harbor off-center mismatches, and SNP prediction is generally not possible. For these regions, hybridization is suppressed for contiguous features in a tiling path. This pattern is therefore a signature of high underlying polymorphism, in the form either of closely linked SNPs or small indels, or potentially of larger deletions (cf. Fig. 1B,C). This phenomenon has limited the utility of resequencing array data for describing patterns of genome-wide sequence variation. Regions where no SNPs are predicted (1) may be monomorphic to the reference sequence or, alternatively, (2) may be so dissimilar that no underlying polymorphisms are detected.

Despite the obvious value in predicting regions of high sequence diversity from resequencing array data, advanced computational approaches to this problem have not been reported. In one study, Hinds et al. (2006) used a simple thresholding algorithm coupled with visual inspection to identify more than a hundred deletions of length 70 bp to 7 kb (median, 750 bp) from resequencing array data for the mouse. More recently, Clark et al. (2007) applied a simple heuristic algorithm to predict tracts of highly divergent or missing sequences from similar data for *Arabi-*

[3]These authors contributed equally to this work.
[4]Corresponding author.
E-mail Gunnar.Raetsch@tuebingen.mpg.de; fax 49-7071-601-801.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.070169.107. Freely available online through the *Genome Research* Open Access option.

18:918–929 ©2008 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/08; www.genome.org

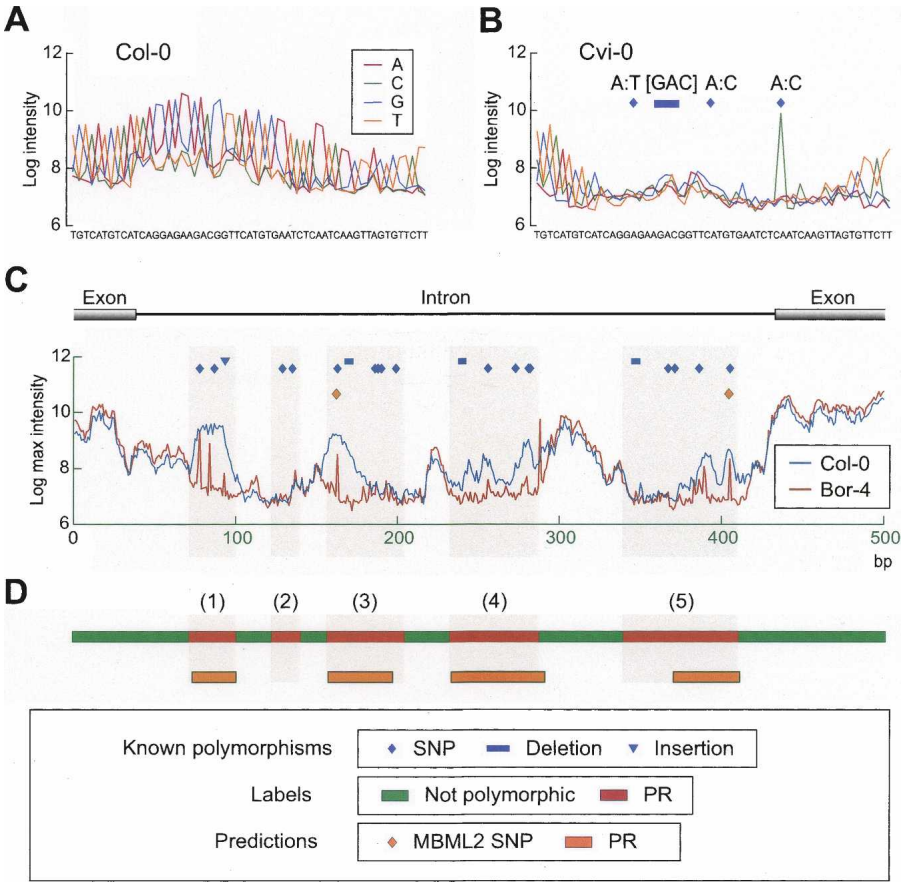

Figure 1. Effect of polymorphisms on hybridization patterns, labels for the mPPR algorithm, and polymorphic predictions. (*A*) Log_2 intensities for oligonucleotides in a 56-bp tiling path (chromosome 4, positions 8,375,747–8,375,802) for the reference Col-0 accession. Intensities for each sequence (see *inset*) are given and are averages for the forward and reverse strand features tiled on the arrays (see Methods). (*B*) Corresponding data from accession Cvi-0 for which three SNPs and a 3-bp deletion are present relative to the tiled Col-0 reference sequence. Intensities are suppressed flanking an isolated SNP (*right*), where the SNP probe shows a clear peak, and intensities for all probes are reduced for the cluster of three polymorphisms, including the deletion (*left center*). (*C*) Log_2 intensities for the maximally hybridizing oligonucleotide at each tiled position are shown for Col-0 and Bor-4 (see *inset*) for a particularly challenging sequence fragment in 2010 (chromosome 3, positions 10,245,203–10,245,702; gene *AT3G27660*). Hybridization properties for much of the region are poor, as reflected by the low intensity values for the perfect match Col-0 reference sequence. Known (2010) and predicted polymorphisms (MBML2) for Bor-4 are as indicated. Only two of the 21 known Bor-4 polymorphisms (17 of which are SNPs) were predicted in MBML2. (*D*) The corresponding polymorphic region (PR) label sequence for Bor-4 and resulting PR predictions (color coding is as shown at *bottom*). Light gray shading that extends across panels *C* and *D* corresponds to PR labels (red). Plotted data are from Nordborg et al. (2005) and Clark et al. (2007).

dopsis thaliana. Although this heuristic algorithm generated several hundred predictions per accession, it only identified extended polymorphic tracts (~300 bp to many kilobases) consisting largely of deletions. Currently, no methods have been reported to predict short indels (tens of base pairs) or clustered SNPs from resequencing array data. This limited investment in methods reflects, in part, the complex nature of the primary data (Clark et al. 2007). In contrast to most microarrays, resequencing arrays harbor all possible oligonucleotides for tiled regions, including those that are repetitive or that have inherently poor hybridization properties. Moreover, replication to reduce experimental noise has typically not been performed for resequencing array studies owing to the high cost of whole-genome analyses (Hinds et al. 2005; Clark et al. 2007; Frazer et al. 2007).

In this work, we describe a machine learning method suitable for predicting regions of high polymorphism density from resequencing array data. Our technique is related to hidden Markov models (HMMs) (e.g., Durbin et al. 1998), which are ubiquitous in computational biology and which have been applied to various segmentation and label sequence learning problems, such as gene finding (e.g., Burge and Karlin 1997). In our case, the prediction task is to label each tiled position in the genome either (1) as conserved or (2) as being at or immediately adjacent to a polymorphism (cf. Fig. 1). The relation between adjoining sites is exploited using a state model representation of these labels (see Methods). For HMMs the label sequence should satisfy the Markov property, and at each time point, observations are assumed to be independent (Durbin et al. 1998). For resequencing array data, the latter assumption is invalid as neighboring 25-mer oligonucleotides overlap and hybridization measurements are highly dependent. Recently, a number of discriminative learning algorithms such as conditional random fields (CRFs) (Lafferty et al. 2001), hidden Markov support vector machines (HMSVMs) (Altun et al. 2003; Tsochantaridis et al. 2005), and the related max-margin Markov networks (Taskar et al. 2003) have been proposed to solve various label sequence learning problems. These methods can handle dependencies between features and have been shown to be very powerful, e.g., for gene finding tasks (Bernal et al. 2007; Rätsch et al. 2007; Schulze et al. 2007). Our method, which we call margin-based prediction of polymorphic regions (mPPR), employs HMSVMs modeling the array measurement sequences to learn to identify polymorphic regions (PRs). Here, we define PRs as contiguous regions of nucleotides, each of which is at most 6 bp from a polymorphism or is between two polymorphisms separated by at most 18 bp (for a discussion of these distances, cf. Supplemental Fig. S1).

We applied mPPR to an *A. thaliana* resequencing array data set for 20 accessions, hereafter called AtAD20, that contains data generated for more than 99.99% of bases in the 119-Mb reference genome (The *Arabidopsis* Genome Initiative 2000) for each accession (Clark et al. 2007). These data were previously used to identify ≈648,000 SNPs at a specificity of ~98% (the MBML2 SNP data set). With mPPR, on average ≈288,000 PRs were predicted per accession at a specificity of ~97%. A large proportion (~66%) of a set of known SNPs were included within PR predictions, of which 42% were absent from the MBML2 data set. The resulting PR data set defines a large fraction of the highly polymorphic or deleted regions segregating in the global *A. thaliana* population, and pro-

vides a high-resolution description of the genome-wide distribution of such regions in a moderately-sized eukaryotic genome.

Results

We adapted HMSVMs to predict PRs from array resequencing data. In brief, HMSVMs try to estimate a function $\boldsymbol{\pi} = f_{\boldsymbol{\theta}}(\mathbf{x})$ of the input sequence $\mathbf{x} = x_1 \ldots x_t$, in our case representing the features derived from the array measurements. This function predicts a label sequence $\boldsymbol{\pi} = \pi_1 \ldots \pi_t$, of the same length t, indicating whether or not a position was within a PR. To estimate the free parameters $\boldsymbol{\theta}$ of function $f_{\boldsymbol{\theta}}$, n training examples, i.e., input sequences $\mathbf{x}^{(i)}$ with corresponding labels $\boldsymbol{\pi}^{(i)}$, $i = 1,\ldots,n$, were used. In its most basic form, the method optimizes the parameters $\boldsymbol{\theta}$ such that there is a large margin between the correct and any incorrect labeling (for details, see Methods), as similarly done in support vector machine classification (e.g., Vapnik 1995; Müller et al. 2001; Schölkopf and Smola 2002).

Our algorithm required a set of accession-matched, known sequences for the generation of label sequences used for training and evaluation. For 19 of the 20 AtAD20 accessions, 1213 fragments of ≈550 bp in length located throughout the genome had been sampled by PCR and dideoxy sequencing (Nordborg et al. 2005). This data set, hereafter called 2010, covers ~0.5% of the genome per accession and harbors ≈2700 SNPs and ≈400 indel polymorphisms per target accession (Nordborg et al. 2005). Col-0, the reference accession, was included in the AtAD20 accession set (Clark et al. 2007), and we used Col-0 array data to assess hybridization performance of arrayed oligonucleotides. As a consequence, predictions could not be generated for Col-0 itself (e.g., to detect errors in the reference sequence) (The *Arabidopsis* Genome Initiative 2000). Our method also used information about the repetitiveness of each arrayed 25-mer oligonucleotide determined from the Col-0 reference sequence (Clark et al. 2007). In particular, we separately modeled repetitive sequences from nonrepetitive sequences in an effort to avoid fragmentation of predictions in regions of low to moderate repeat content (see Methods).

Performance evaluation on 2010

We trained our method on 60% of the 2010 data, and used 20% for hyper-parameter tuning and 20% for evaluation; we employed a fivefold cross-validation strategy to obtain out-of-sample predictions for all 2010 fragments. For our method, we considered a prediction as a true positive (TP) if a portion λ (or more) was covered by PR(s); else it was counted as a false positive (FP). Conversely, a known PR was counted as a true discovery (TD) if all underlying polymorphisms were inclusive to a prediction or if at least λ of its length was contained in one or more PR prediction(s); else it was a false negative (FN). We used these counts to assess specificity and sensitivity (for details, see Supplemental Fig. S2 and Methods), and we excluded PRs from evaluation that were more than 75% duplicated elsewhere in the reference genome (these repetitive PRs constituted 3.4% of examples in 2010).

Tuning an internal parameter of our algorithm on the five cross-validation sets allowed us to adjust the trade-off between specificity and sensitivity (for details, see Fig. 2A and Methods). For 2010, we generated predictions at a specificity of ≥90% for λ = 75% (for the effect of varying λ on specificity and sensitivity, see Supplemental Fig. S3). Across all sequence types and accessions, our method identified 56% of PRs in 2010, and performance estimates varied only moderately between accessions (Supplemental Table S1). In *A. thaliana*, coding sequences have higher GC content and sequence complexity than noncoding sequences (The *Arabidopsis* Genome Initiative 2000). These factors are favorable for hybridization-based methods (Lee et al. 2004; Clark et al. 2007) and likely contributed to the higher sensitivity in coding regions (e.g., about a 1.3-fold difference compared with noncoding sequences at a similar specificity; cf. Fig. 2A; Table 1). As minor differences in prediction boundaries affect performance estimates—especially for small predictions (cf. Supplemental Fig. S2)—we also assessed the performance of the predictions with a relaxed overlap criterion. For λ = 50%, sensitivity was slightly higher, and specificity was at least 95% for all sequence types and ≈97% on average (cf. Table 1).

The labels we used for training are abstractions for underlying polymorphism; however, all polymorphism types were labeled (e.g., both SNPs and indels) and were thus targets for prediction. We therefore assessed the polymorphism content of predictions on the 2010 test data. Sixty-two percent of predictions identified single SNPs, 3.4% harbored single indels, and the remaining predictions identified complex mixes of polymorphism types, with clusters of SNPs most common (Supplemental Table

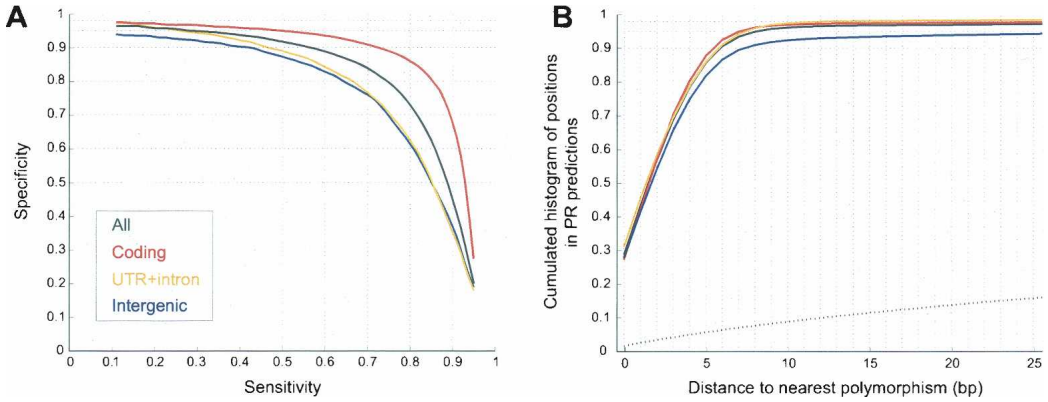

Figure 2. Relationship between specificity and sensitivity for PR predictions with overlap criteria λ = 75%. (*A*) Specificity–sensitivity curves averaged over cross-validation test subsets for different sequence types (for color code, see *inset*). PRs that contained more than one sequence type were assigned to the type comprising the majority of the prediction. (*B*) Specificity at the nucleotide level as calculated for each position within a prediction. Deleted nucleotides and SNP positions were assigned a distance of 0. A cumulative histogram of these distances is displayed, showing that, e.g., more than 90% of all nucleotides in PR predictions are within six nucleotides to a known polymorphism. The dotted black line indicates the relationship expected by chance (i.e., predictions were assigned to random genomic locations for calculating distances).

Table 1. Specificity and sensitivity for PR predictions assessed with 2010 for different overlap cut-offs, λ (see main text)

	Coding	UTRs + introns	Intergenic	2010	Genome
λ = 75%					
Specificity	92.6%	88.8%	88.3%	90.4%	89.3%
Sensitivity	63.6%	50.9%	49.1%	55.6%	52.0%
λ = 50%					
Specificity	97.4%	97.9%	95.8%	97.2%	96.6%
Sensitivity	65.5%	54.4%	51.7%	58.2%	54.7%

The relative abundance of sequence types differs between 2010 and the whole genome (Clark et al. 2007), and specificity and sensitivity were re-estimated accordingly for the whole-genome predictions (Genome column).

S2). For indel polymorphisms, 53.3% of deleted bases and 38.9% of insertion sites in 2010 were included within predicted PRs. Across all prediction types, ~90% of bases within predictions were at or within 6 bp to a known polymorphism (see Fig. 2B).

While PR predictions typically reflected the underlying patterns of polymorphisms with high accuracy, prediction boundaries sometimes differed substantially from labels, and for some regions, even highly clustered polymorphisms were not identified (Fig. 1C,D). In large part, such FNs occurred for regions with poor hybridization properties in the reference accession (e.g., Fig. 1C,D, cf. predictions to reference feature intensities for regions 2 and 5; see also Supplemental Fig. S4). Additionally, although explicitly modeled by our method, repeats were overrepresented among FN predictions. For example, in 2010, 5.5% of all positions were repetitive (see Methods), while the fraction of repetitive positions in FN PRs was twice as high (10.9%). In contrast, only 2.1% of sites in correctly predicted PRs were repetitive. Therefore, repeats are a source of error for our predictions; however, mPPR was cautious in making predictions that included repetitive sites.

Prediction content and comparison to MBML2

We designed mPPR to produce predictions that complement existing SNP data sets ascertained from resequencing array data (Fig. 3). Although our method only identifies the approximate location of polymorphisms, 74.8% of clustered SNPs (≤18 bp away from the nearest polymorphism) in 2010 were included within boundaries of PR predictions (Table 2). This contrasts markedly to MBML2, for which a mere 12.4% of the clustered SNPs were identified. Although mPPR performed well for clustered SNPs, the method nevertheless also identified 55.4% of isolated SNPs (those >18 bp to the nearest polymorphism). Compared with MBML2, 42% of 2010 SNPs were located exclusively within mPPR prediction boundaries, whereas only 8% were found exclusively in MBML2. The most striking differences between the data sets were for clustered SNPs in untranslated and intergenic regions, where our method identified the approximate location of sevenfold to 10-fold as many SNPs as MBML2.

Whole-genome predictions and evaluation

HMSVMs trained on 2010 data were used for genome-wide prediction on AtAD20 accessions using the same settings as for evaluations on 2010 data (Table 1). Nonredundantly, 27% of the *A. thaliana* genome was included within the boundaries of the resulting predictions, and 92% of the predictions harbored <75% repetitive sites, the criteria we used for evaluation with 2010. Per accession, between 240,538 and 361,184 PRs were predicted, comprising between 5.3% and 8.5% of the genome (Supplemental Table S1). The accession with the most predictions, Cvi-0, was known from earlier work to be highly dissimilar to Col-0 (Schmid et al. 2003; Nordborg et al. 2005). By sequence type, intergenic positions were most strongly overrepresented within prediction boundaries (Supplemental Fig. S5).

Given the size and genome-wide sampling for the 2010 data (Nordborg et al. 2005), our performance evaluations likely generalize well for much of the genome. Nevertheless, the 2010 data are biased in several ways that potentially affect performance estimates. First, 2010 is overrepresented for coding sequences, and we adjusted performance estimates for genome predictions to account for the difference in sequence composition between 2010 and the whole genome (Table 1). However, noncoding sequences in 2010 are also biased, and are generally located in close proximity to coding sequences. A consequence is that polymorphism levels for the 2010 sequences are likely reduced compared to the genome average. Another concern is that, irrespective of sequence type, the PCR-based 2010 data are underrepresented for highly divergent or deleted sequences that could not be amplified by PCR.

We therefore used several resources partially or entirely independent of 2010 to evaluate genome-wide predictions. First, we assessed prediction quality using clone-based genomic sequence data available for three of the studied accessions. This included 37 kb of BAC sequences available for accession Cvi-0 and 14 kb for C24. Here, specificity was 96% and 100% (for λ = 50%) at a sensitivity of 67% and 45% for Cvi-0 and C24, respectively (Supplemental Table S4). Moreover, we assessed our predictions using the much larger twofold draft shotgun sequence data available for Ler-1 (see Supplemental Methods). Although we excluded repetitive regions from this evaluation, performance estimates with this genome-wide resource are expected to be largely unbiased by sequence composition. After removing contigs that were likely the result of assembly errors (see Supplemental Methods), the prediction quality assessed with 37.9 Mb of aligned sequence data differed only marginally from that assessed with the 2010 test data (e.g., specificity was 96%) (Supplemental Table S4). Thus, performance estimates with the genomic clone data were in general agreement with the PCR-based test data even though the composition of the predictions differed somewhat from those in the 2010 test set (e.g., more PRs harbored clusters of SNPs or indels; Supplemental Table S2).

Second, we assessed the performance of predictions for long deletions, a polymorphism type absent from 2010 and that we excluded from the clone-based data owing to alignment uncer-

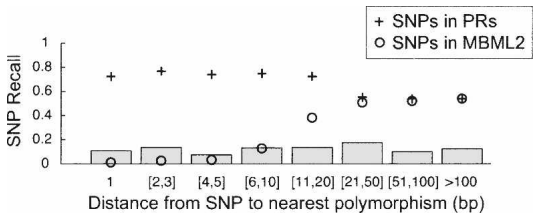

Figure 3. Dependency of SNP sensitivity on distance between polymorphisms by detection method. SNPs were partitioned according to the distance to the nearest polymorphism. The frequency of SNPs in each distance bin (*X*-axis) is shown as bars. Sensitivity rates per distance category are given for MBML2 SNP calls (circles) and inclusion within PR prediction boundaries (crosses).

Table 2. Sensitivity by polymorphism and sequence type

	Coding		UTR + intron		Intergenic		All	
	PRs	MBML2	PRs	MBML2	PRs	MBML2	PRs	MBML2
Clustered SNPs	85 (65) [9196]	21	71 (64) [10,793]	7	66 (57) [5608]	9	75 (63) [25,597]	12
Isolated SNPs	61 (14) [10,294]	69	54 (22) [6774]	41	48 (22) [5870]	37	55 (18) [22,938]	53
All SNPs	72 (38) [19,490]	46	64 (48) [17,567]	20	57 (39) [11,478]	23	66 (42) [48,535]	31

For MBML2 (specificity ≈ 98%; Clark et al. 2007), the percentage of SNPs for which the correct position and allele was identified is given; for the PR data, the percentage of SNPs contained within PR prediction boundaries is given (specificity ≈ 90%; cf. Table 1), with the percentage of SNPs contained within PR predictions but absent from MBML2 given in parentheses. SNPs were classified as isolated if the distance to the nearest polymorphism was >18 bp, otherwise as clustered. Sample sizes are indicated in brackets. Untranslated regions (UTRs) and introns were evaluated together owing to small UTR sample size in 2010.

tainties in the draft genomic data (see Supplemental Methods). More than 100 known deletions of greater than 300 bp had been previously characterized in AtAD20 accessions (Clark et al. 2007) or were characterized in the current study (see Supplemental Methods). These deletions were almost entirely included within PR predictions (Fig. 4B; Supplemental Table S5; Supplemental Fig. S6).

Finally, we note that extended tracts of repetitive sequences (>500 bp) are entirely absent from our evaluations (see Supplemental Methods). Nonetheless, such sequences are common in *A. thaliana* and are dispersed throughout the genome. To evaluate these as potential sources for false predictions, we took advantage of large regions known to be substantially identical to the Col-0 reference. Previously, Toomajian et al. (2006) used 2010 data to infer regions of extended haplotype sharing (i.e., sequence identity) with the Col-0 genome for the AtAD20 accessions. In such accessions and regions, our method predicted few PRs, e.g., as can be seen for a 600-kb region in Est-1 for which all 2010 segments are identical to Col-0 (Fig. 4A; Toomajian et al. 2006; Clark et al. 2007). This suggests a low incidence of false predictions in regions that are monomorphic to the reference genome sequence but that have repetitive sequence compositions broadly representative of the *A. thaliana* euchromatic genome.

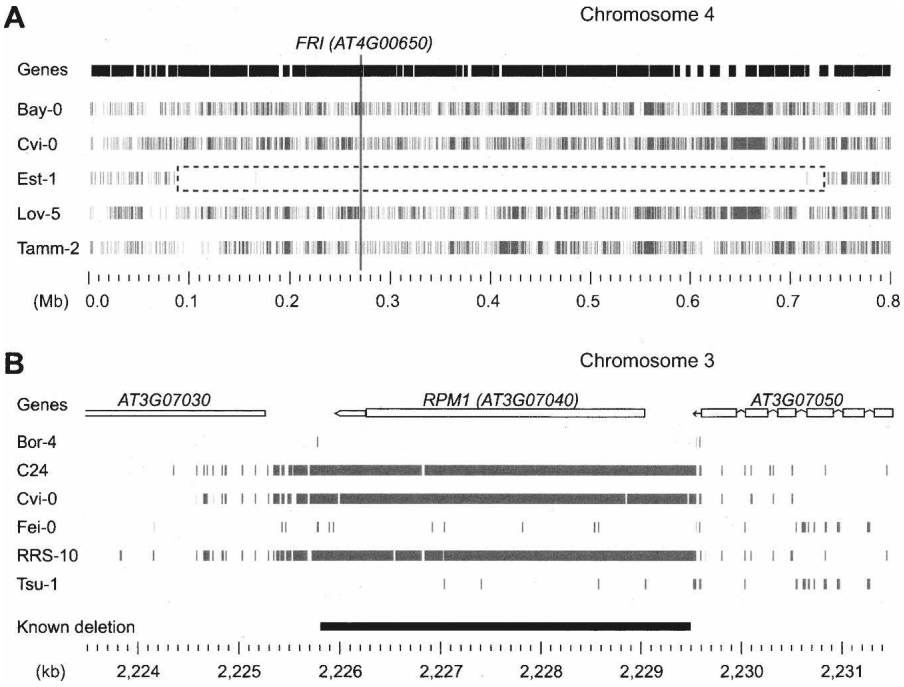

Figure 4. PRs reveal haplotype sharing at chromosomal and local scales. (*A*) Genes (*top*) and PRs (gray blocks beneath) for five accessions for 0.8 Mb surrounding the *FRI* locus. In Est-1 a region of ~0.6 Mb (dashed black box) including *FRI* (vertical line) has been reported to be nearly identical to the Col-0 reference sequence but divergent in the other accessions shown (Nordborg et al. 2005; Clark et al. 2007). Only several PRs are located in the Est-1 region that is monomorphic with the tiled reference sequence. (*B*) Pattern of PRs for 8 kb at the *RPM1* locus. The location of a 3.7-kb deletion that segregates in the *A. thaliana* population is as indicated at *bottom* (Grant et al. 1995, Shen et al. 2006). Experimental characterization revealed that the C24, Cvi-0, and RRS-10 accessions included in the current study harbored this deletion (the other accessions shown have a Col-0 like haplotype). PRs delineate the deletion as well as flanking SNPs and indels (see also Supplemental Fig. S6).

Polymorphism patterns ascertained with PR and MBML2 data

An immediate use of PR predictions is the characterization of genome-wide patterns of genetic variation. While PR predictions delineate clusters of SNPs and indels with high accuracy, the nature of polymorphism underlying a given prediction is unknown. To examine genome-wide polymorphism levels, we therefore simply counted whether a base was included in a PR prediction in one or more of the AtAD20 accessions. To provide insights into ascertainment biases introduced by different methods, we also calculated the analogous polymorphism estimate with MBML2 SNP data.

Despite the inherent differences in prediction methods, patterns of polymorphism assessed using the PR and SNP data sets were nonetheless broadly correlated at chromosomal scales (Fig. 5; Supplemental Fig. S7). Polymorphism patterns apparent in the PR data also resembled that for pairwise nucleotide diversity, as previously calculated with MBML2 (Clark et al. 2007), as well as for several data sets generated by dideoxy sequencing (Nordborg et al. 2005; Schmid et al. 2005; Clark et al. 2007). Moreover, the patterns were also similar to those observed in single feature polymorphism data collected with the *A. thaliana* ATH1

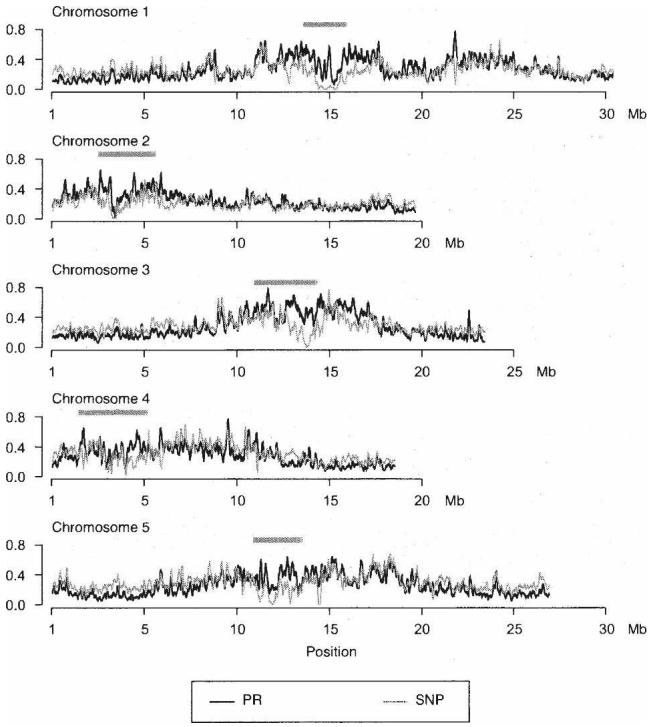

Figure 5. Genome-wide patterns of polymorphism in PRs and MBML2 SNPs. A sliding window of 100 kb was used, with values for every 10,000th position plotted. The *Y*-axis displays the fraction of bp in each window included within PRs nonredundantly over all accessions (black line), and the two measures of polymorphism are broadly correlated (Supplemental Fig. S7). To facilitate visualization, the analogous measure for the SNP data was multiplied by 50 (gray line). Thick gray bars indicate the approximate positions of centromeres as defined by repeat content in an earlier study (Clark et al. 2007).

microarray (Borevitz et al. 2007). In particular, polymorphism tended to be higher for centromeric and pericentromeric sequences, with additional regions of extended high polymorphism also apparent on chromosomal arms (e.g., distal to the centromeres on chromosomes 1 and 5) (Fig. 5).

We also examined polymorphism levels by sequence type by determining, for each position, the fraction of bases included in predictions across all accessions. Here, polymorphism apparent in PR and SNP data varied in a manner consistent with ascertainment biases (Table 2; Clark et al. 2007). Within genes, predicted polymorphism levels were on average higher for intronic sequences than for coding sequences when assessed with PR, but not with MBML2 data (Fig. 6A). For the PR data, the observed pattern is consistent with the general expectation of reduced evolutionary constraint for nontranslated sequences, as well as with estimates of nucleotide diversity from 2010 (Nordborg et al. 2005). In addition, inclusion of indels as prediction targets for mPPR, coupled with the bias for indel polymorphisms in noncoding regions (The *Arabidopsis* Genome Initiative 2000), is a likely factor contributing to fine-scale differences in polymorphism estimated from the different data sets.

We also used PR data to infer the distribution of polymorphisms in intergenic sequences for which SNP sensitivity for MBML2 is very low (Table 2; Clark et al. 2007) and for which diversity estimates from 2010 are largely limited to sequences near genes (Nordborg et al. 2005). Average levels of polymorphism varied as a function of distance from coding sequences and were asymmetric relative to gene orientation (Fig. 6B). Extending upstream to 5′ UTRs, polymorphism reached a plateau at ~450 bp, while the analogous plateau was reached within ~50 bp downstream from 3′ UTRs. Upstream of transcription start sites, polymorphism tended to be inversely associated with the density of predicted *cis*-regulatory elements (O'Connor et al. 2005). The reduced polymorphism 5′ to genes may, therefore, reflect constraint on *cis*-regulatory sequences, as suggested by permutation tests that revealed a highly significant under-representation for PR overlaps to predicted *cis*-regulatory sites (Fig. 6C; Supplemental Fig. S8; O'Connor et al. 2005).

Highly polymorphic genes and gene families in *A. thaliana*

At the local scale, we used PR predictions to characterize, at high resolution, genes that are highly polymorphic in the *A. thaliana* population. On an accession basis, an average of 117 of 26,541 coding genes had more than 75% of their coding sequence within predictions. Across all accessions, we also assessed patterns of polymorphism among classes of genes by determining the fraction of coding bases per gene included in PR predictions (denoted "PR content"). Globally, intraspecific patterns of genic polymorphism predicted interspecific conservation, with lower PR content for *A. thaliana* genes with orthologs in black cottonwood (*Populus trichocarpa*), the most closely related plant with a sequenced genome (Supplemental Fig. S9; Tuskan et al. 2006). Among large gene families within *A. thaliana* ($n > 125$) (Clark et al. 2007), variation in PR content was readily apparent (Fig. 7A,B; Supplemental Fig. S10). Transcription factors, for which MBML2 SNP data suggested strong purifying selection, harbored few members with high PR content (Supplemental Fig. S10). In contrast, higher PR content was observed for F-box genes (Supplemental Fig. S10), for which many inactivating mutations have been identified (Clark et al. 2007), and for which patterns of sequence variation indicate high death rates in the *A. thaliana* genome (Thomas 2006). Among large gene families, nucleotide-binding leucine rich repeat (NB-LRR) genes that mediate disease resistance harbored extreme levels of polymorphism (Fig. 7B), a finding that was even apparent in low resolution predictions of PRs from AtAD20 data (Clark et al. 2007).

As our PR predictions have high specificity and sensitivity in noncoding regions, we also used PR content to assess sequence variation within and among micro-RNA (miRNA) genes, where comparatively little is known about within-species polymorphism. Among *A. thaliana* miRNAs with homologs in other species (Jones-Rhoades et al. 2006), very little variation was observed for the 21-nucleotide miRNA sequences required for miRNA mediated gene suppression (Fig. 7C). Marginally higher variation was observed for the complementary miRNA* sequence, with PR content substantially higher for precursor end and loop regions of miRNA precursor sequences. For a set of 68 validated or predicted miRNAs lacking homologs in other species (Rajagopalan et al. 2006; Fahlgren et al. 2007), PR content was generally much higher, and the pattern of reduced PR content for the miRNA sequence relative to the rest of the precursor was less clear (Fig. 7C). Whether this pattern reflects poor annotation for the nonconserved miRNAs, or potentially the evolution of new genes that are not fixed in the population, remains to be determined.

Data release

The PR prediction data set is available for download from The *Arabidopsis* Information Resource (TAIR; http://www.

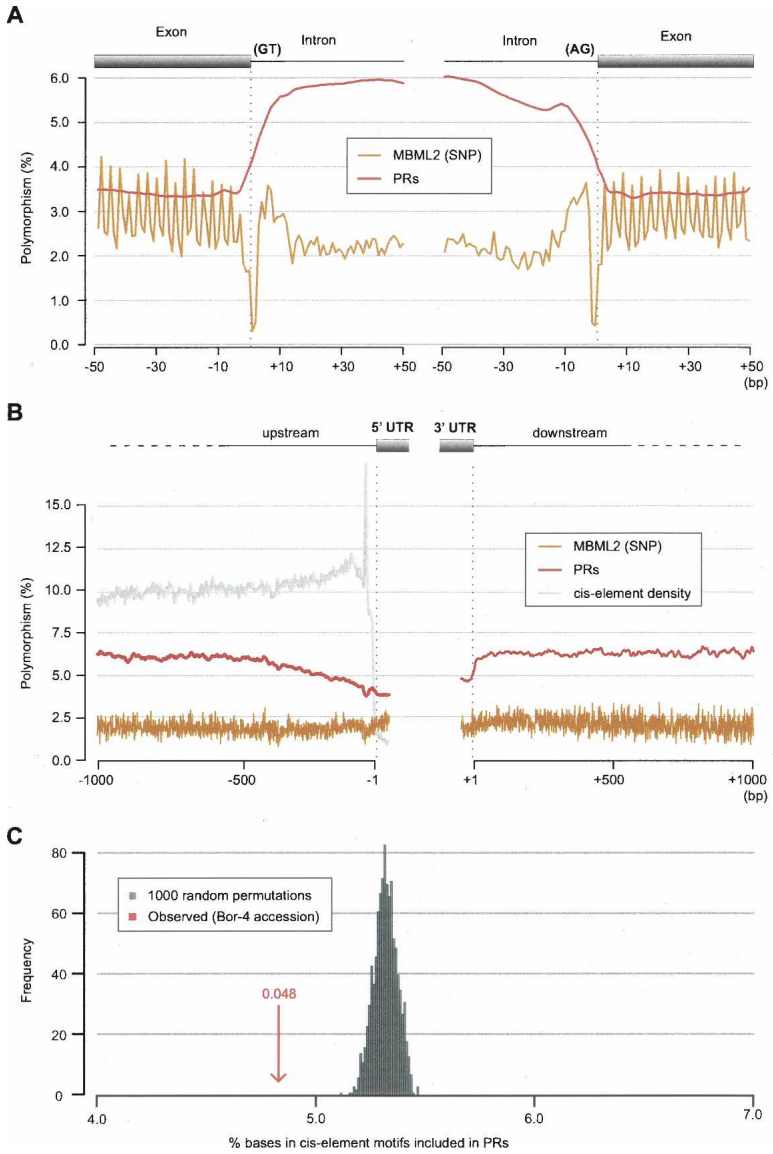

Figure 6. Patterns of polymorphism apparent in PR and SNP data in noncoding regions. (*A*) Polymorphism near splice donor (*left*) and splice acceptor (*right*) sites as averaged over 116,971 splice sites and assessed with both the PR prediction and MBML2 (SNP) data sets (for details of polymorphism estimation, see *inset*; Supplemental Methods). Relaxed constraint at wobble positions is apparent in the SNP data as sequential peaks in polymorphism with a 3-bp offset (the observed pattern reflects, in part, biased splicing at codon boundaries). SNP polymorphism is lowest at splice sites, and polymorphism estimates with the PR and SNP data diverge for intronic sequences (*middle*). (*B*) Comparison of the PR and SNP polymorphism estimates for the 1000 bp located 5′ and 3′ to transcription units for coding genes (averaged across 17,434 genes with annotated 5′ UTRs, and 17,430 genes with annotated 3′ UTRs). The average density of predicted *cis*-elements for the 5′ region is as shown. A peak immediately 5′ to transcription start sites corresponds to the TATA motif. (*C*) Percentage overlap of PRs to *cis*-element motifs mapped to the *A. thaliana* genome for 9599 upstream regions for Bor-4 (red arrow) (for overlap in other accessions, see Supplemental Fig. S8). The overlap expected by chance was established by permuting PRs and upstream regions 1000 times (gray shading; see Supplemental Methods).

arabidopsis.org), as are lists for the percentage of each coding gene included in predictions by accession.

Discussion

Because array-based resequencing relies on hybridization, highly polymorphic regions present a substantial challenge. Generally, few SNPs are predicted for these regions, and precise methods for prediction of indels have not been developed. Nevertheless, clustered polymorphisms and indels, which can comprise more than 15% of polymorphisms in eukaryotic genomes (e.g., Dawson et al. 2001; Wicks et al. 2001; Mills et al. 2006), are a central component of sequence variation and contribute to phenotypic variation. Here, we present a method, mPPR, for accurate prediction of PRs from resequencing array data of the reference plant *A. thaliana*, where SNP polymorphism is higher than for human (Wright and Gaut 2005 and references therein), and for which indel polymorphisms are common (The *Arabidopsis* Genome Initiative 2000; Nordborg et al. 2005). While replicated hybridization measurements are typically not available for primary whole-genome hybridization data, each base in a tiling path is interrogated on the arrays, an ultimate determinant for the theoretical accuracy of predictions. By using a machine learning method to overcome experimental noise and to relate complex, dependent hybridization measurements from overlapping oligonucleotides to underlying polymorphisms, we detected even small clusters of SNPs or indels (within less than 10 bp) with high accuracy. A challenge for our learning method were large deletions that were absent from the 2010 training data. Nonetheless, deletions maximally suppress intensity measurements throughout a tiling path, and suppressed hybridization is the pattern identified by mPPR. In our predictions, long deletions were readily recognizable as (potentially interrupted) long PRs (for an example, see Fig. 4B).

Fine-scale patterns of polymorphisms

Our PR predictions, which are genome-wide and largely unbiased by sequence type, revealed patterns of polymorphism not apparent in earlier analyses. For example, in intergenic sequences, we found that average polymorphism is lowest immediately 5′ to transcribed sequences, rising to maximal levels within ~450 bp of the transcription start site. This observation is unlikely to result from an artifact in the PR data; a similar pattern is apparent in an interspecific comparison of promoter regions between *A. thaliana* and a close relative, *Boechera stricta* (Windsor et al. 2006). Constrained sequence evolution for regions immediately 5′ to genes may reflect the action of purifying selection on *cis*-regulatory sequences, as suggested by a significant underrepresentation of overlaps between PRs and

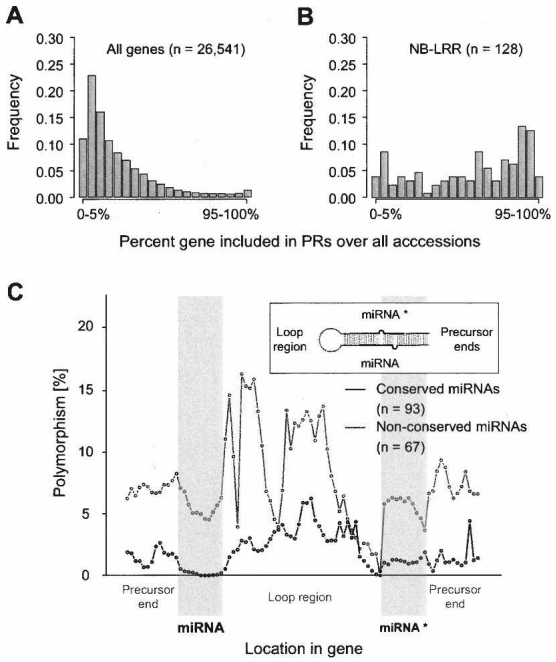

Figure 7. Percentage of coding and miRNA genes included in PRs over all accessions by gene category. (*A*, *B*) Distribution of coding genes as a function of percentage inclusion in PRs for all genes and NB-LRR genes, respectively (see Supplemental Methods). (*C*) Polymorphism averaged over conserved and nonconserved miRNA genes by location in the stem–loop structure (*inset* and as labeled at *bottom*). To facilitate visualization, lengths of the stem-loops were scaled relative to each other as described in the Methods.

transcriptional *cis*-elements predicted in a previous study (O'Connor et al. 2005). This finding suggests that in *A. thaliana*, the information required for gene expression is densest in close proximity to transcript start sites even though full recapitulation of complex expression patterns often requires substantially larger promoter fragments (e.g., Lee et al. 2006). An implication of this observation is that deep sampling of variation within *A. thaliana* populations will be important for both detecting *cis*-regulatory sequences and characterizing their evolution.

The specificity of our predictions also allowed us to characterize polymorphisms in transcribed *A. thaliana* sequences at a resolution of tens of base pairs. Hundreds of transcribed regions, representing genes from many families, were largely covered by PRs in one or more accession. In some cases, this may reflect the absence of selection at annotated genes that are in fact pseudogenes. In other cases, highly dissimilar sequences may reflect the action of balancing selection, where linked mutations accumulate nearby a selectively maintained polymorphism. Allele frequency patterns in SNP data support balancing selection as a central force leading to high polymorphism levels for NB-LRR genes (Bakker et al. 2006; Clark et al. 2007), the predominant class of disease resistance (R) genes in plants (Jones and Dangl 2006). In our study, family-wide polymorphism for NB-LRR genes was extreme, as also noted from earlier work with the AtAD20 data (Clark et al. 2007), as well as from studies of a select set of NB-LRR genes in *A. thaliana* (Grant et al. 1998; Bakker et al. 2006; Shen et al. 2006). Nevertheless, polymorphism levels for individual NB-LRR genes varied greatly; some genes were almost entirely included in PRs, while others were predicted to be largely monomorphic across the AtAD20 accession set. This might reflect the action of different selective pressures on specific family members, and NB-LRR genes harboring little or no variation may have been targets of recent positive selection (sweeps) in *A. thaliana* populations. Although the primary function for NB-LRR genes is in race-specific resistance to pathogens, not all R genes are NB-LRR members (e.g., Song et al. 1995). The extent to which other highly polymorphic genes identified in this study mediate interactions with the biotic (or potentially abiotic) environment requires empirical study.

Utility of predictions for functional studies

Our predictions are immediately useful for functional studies in *A. thaliana*. Many genes entirely covered by PRs are likely to be partially or completely deleted. These constitute a potential source of loss-of-function alleles for genes for which knockout alleles have not been found in sequence indexed *A. thaliana* mutant collections (Alonso and Ecker 2006). Moreover, the AtAD20 set was selected not only to maximally capture diversity within the species but also to include many parents of recombinant inbred line (RIL) populations constructed for quantitative trait locus (QTL) mapping (http://www.inra.fr/internet/Produits/vast/RILs.htm). Deletions or highly polymorphic sequences have been shown to underlie diverse phenotypes that segregate in *A. thaliana* populations (e.g., Johanson et al. 2000), and our predictions should be valuable for identifying causal alleles found in QTL studies, or that are linked to SNPs employed in whole-genome association mapping scans (Kim et al. 2007). At a more basic level, our predictions will facilitate the design of perfect match primers for genotyping and for collecting diversity data with PCR-based methods. Further, the predictions are useful for identifying mismatched probes present on microarrays employed for interrogating RNA expression in different accessions.

Application of our methods to other data and broader relevance

Although mPPR was tailored for predicting PRs with *A. thaliana* resequencing array data, it should be readily applicable to other resequencing array data sets with some modifications. In previous experiments with human and mouse (e.g., Hinds et al. 2005; Frazer et al. 2007), and for ongoing work with rice (McNally et al. 2006), DNA hybridized to arrays was generated by pooling long-range PCR amplicons of selected regions. For *A. thaliana*, the entire genomic DNA was subjected to isothermal amplification (Clark et al. 2007). Nevertheless, the framework of our learning algorithm can be adapted to accommodate additional intensity variation resulting from concentration differences between individual long-range PCR products. In humans, heterozygosity presents an additional challenge, as does the lack of sample-matched training data. In contrast, for other species, hybridization was performed using inbred (homozygous) strains (Frazer et al. 2004, 2007; McNally et al. 2006), and sample-matched data sets that could potentially be used for training have been reported for mouse (e.g., Mural et al. 2002) and are being generated for rice (K. Childs and R. Buell, pers. comm.) that would facilitate application of the mPPR approach.

In this work we adopted a new machine learning method for inferring structural information from sequences to the problem of identifying polymorphisms. The underlying inference method has recently been developed in the field of machine learning (Altun et al. 2003; Tsochantaridis et al. 2005; Rätsch and Sonnenburg 2007) and can be seen as an extension of support vector

machines that are frequently used in computational biology. It has been successfully applied in natural language processing (e.g., Altun et al. 2003; Sha and Saul 2007), computational gene finding (Rätsch et al. 2007), and spliced sequence alignment (Schulze et al. 2007). This diversity illustrates the flexibility and power of the approach. Traditionally, generative models such as HMMs have been used for similar applications. They attempt to estimate probability densities over the observation sequence and their segmentation. However, it has been argued that such approaches do not lead to the best discrimination performance, as high-dimensional density estimation is known to be a harder task than discrimination (Vapnik 1995). One reason generative methods are often outperformed by discriminative methods is that they typically need to assume independence between observations in a sequence (for a comparison of label sequence learning methods and discussion, see also Nguyen and Guo 2007). Since our method does not assume independence, it is well-suited for many tasks in genome research for which measurements are dependent.

Methods

Preparation of hybridization, repeat, and sequence data

Hybridization data from Clark et al. (2007) were quantile-normalized (Bolstad et al. 2003) to correct for between-array variation in hybridization intensities and to facilitate the use of predictors trained with data from all accessions. Consequently, predictors were available to make predictions on any accession. The 2010 data set that we used to generate the label set for both PRs and conserved regions (see below) is available for download as previously described (Nordborg et al. 2005; Clark et al. 2007). Array measurements for repetitive oligonucleotides are much less reliable than for unique oligonucleotides; therefore, we annotated repetitive 25-mer oligonucleotides on the resequencing arrays as described by Clark et al. (2007). We combined information for all types of 25-mer repeats defined by Clark et al. (2007) to create a 0/1-sequence that indicated whether a site was repetitive according to any of the categories. This repeat-mask (called RM) was an input for our algorithm.

Overview of the mPPR algorithm

We started by introducing a graphical model of states and allowed transitions (Supplemental Fig. S11). Instead of predicting the label (polymorphic or conserved) directly, our algorithm was designed to learn to assign a state to each sequence position given the hybridization measurements. To do this, each known sequence in the 2010 data set was first translated into a state sequence, i.e., the "truth" that we tried to approximate. We then applied HMSVMs (Altun et al. 2003) for label sequence learning. We augmented these with explicit feature scoring functions and adapted these to our task by defining an appropriate loss function. From the predicted state sequence, we afterward inferred the label sequence (see color coding in Supplemental Fig. S11).

State model

The simplest possible model, with one state, C, for conserved nucleotides and one state P for PRs, was extended in two ways. First, we noted that hybridization signal gradually decreases over a few nucleotides toward a polymorphism. We therefore included a series of three states—T_1, T_2, T_3—modeling decreasing intensities upstream of PRs and similarly three states—T_4, T_5, T_6—for increasing intensities downstream (for details, see Supplemental Fig. S11). The second extension relates to repetitive sequences, which we modeled separately from unique sequences via duplicated states that effectively allowed feature scoring functions for repetitive regions to be learned differently. The model contains a state C_R for conserved, repetitive sequences (positions p where $RM(p) = 1$) and a state C_U for conserved, unique sequences (where $RM(p) = 0$), likewise a state P_U for polymorphic, unique sequences with P_R as the repetitive counterpart. Transition states T_i were not duplicated. We denote the set of states by S. Allowed transitions between the states are drawn as arcs in Supplemental Figure S11. Real-valued scores $\phi(i,j)$ were associated with transitions from state $i \in S$ to state $j \in S$, which were determined during training of the method except for the transitions $\phi(i,c_R)$ or $\phi(i,c_U)$ that were made deterministically depending on whether $RM(p) = 1$ or $RM(p) = 0$, respectively (and similarly for $\phi(i,P_R)$ and $\phi(i,P_U)$).

Generation of labels

To train our method, we first generated the target state sequence that is to be reproduced given only the input sequence. Initially, all polymorphic sites (deleted nucleotides, SNPs, and nucleotides directly upstream of an insertion site) known from the 2010 set were assigned P_U or P_R states depending on the repeat annotation. In the next step, we assigned P_U or P_R states to sites between two polymorphic labels at a distance of $\leq$18 bp (for the choice of this distance, cf. Supplemental Fig. S1). Every segment of P states was then extended 6 bp in each direction, and the transition states $T_1,\ldots,T_3$ and $T_4,\ldots,T_6$ were inserted upstream and downstream of every segment of P states, respectively. Finally, C_U or C_R states were assigned to the remaining positions. This procedure generated a state sequence for every fragment in the 2010 data set.

Generation of input features

As input to our learning algorithm, seven features were derived from hybridization data. Some of these also used information from the reference genome sequence. Three groups of features were used. First were features directly derived from array intensities (cf. Supplemental Table S6, features 1–4). Some of these were based on a ratio between hybridization intensities of the target and the reference accession. Second, one feature was computed from quality scores (feature 5). Third, several features were included that capture the (dis)agreement between raw base calls from the arrays and the reference sequence (features 6, 7). Quality scores and raw base calls were as defined previously (Clark et al. 2007). The result was a feature vector of length $m = 7$ associated with every position in the genome. Additionally, the repeat annotation *RM* was included; however, this was used to switch deterministically between C_U and C_R states, as well as between P_U to P_R, and not for learning per se.

Parametrization

Formally, our goal was to learn a function

$$f: X \rightarrow S^*$$

that predicts the state sequence (path) $\boldsymbol{\pi} \in S^*$ given the sequence of observations $\mathbf{x} \in X$ (input features), both of equal length t, where S^* denotes the Kleene closure. This was done indirectly via a $\boldsymbol{\theta}$-parametrized discriminant function

$$F_{\boldsymbol{\theta}}: \mathrm{X} \times S^* \rightarrow \mathbb{R}$$

that assigned a real-valued score to a pair of observation and state sequence (Altun et al. 2003). Once $F_{\boldsymbol{\theta}}$ is known, f can be obtained as

$$f(\mathbf{x}) = \underset{\boldsymbol{\pi} \in S^*}{\operatorname{argmax}} F_{\boldsymbol{\theta}}(\mathbf{x},\boldsymbol{\pi}).$$

In our case $F_{\boldsymbol{\theta}}$ satisfied the Markov property, which is sufficient to show that this decoding can be computed efficiently by dynamic programming (Durbin et al. 1998; Giegerich et al. 2004).

The input to the discriminant function $F_{\boldsymbol{\theta}}$ consisted of observations $\mathbf{x}$, an $m \times t$ matrix of m different features, and a sequence of states $\boldsymbol{\pi} = \pi_1,\ldots,\pi_t$. For every pair of features $j = 1,\ldots,m$ and states $k \in S$, we employed a feature scoring function $g_{j,k}$: $\mathbb{R} \to \mathbb{R}$. $F_{\boldsymbol{\theta}}$ was then obtained as a linear combination of the feature scoring contributions and the transition scores ϕ:

$$F_{\boldsymbol{\theta}}(\mathbf{x},\boldsymbol{\pi}) = \sum_{p=1}^{t}\sum_{j=1}^{m}\sum_{k\in S}[[\pi_p = k]]\, g_{j,k}(x_{j,p}) + \phi(\pi_{p-1},\pi_p),$$

where [[.]] denotes the indicator function. For convenience of notation, we assumed a pseudo-transition $\phi(\pi_0,\pi_1) = 0$. We modeled the feature scoring functions $g_{j,k}$ as piecewise linear functions as follows (Rätsch et al. 2007): Let Q be the number of supporting points q_l (satisfying $q_l < q_{l+1}$) and v_l their values, then the piecewise linear function is defined by

$$g(x) = \begin{cases} v_1 & x < q_1 \\ \dfrac{v_l(q_{l+1}-x) + v_{l+1}(x-q_l)}{x_{l+1}-x_l} & q_l \le x < q_{l+1}. \\ v_Q & x \ge x_Q \end{cases}$$

We chose $Q = 10$ supporting points on the abscissa such that in each interval $[q_l, q_{l+1}]$ there were approximately equally many feature values (determined on the training set). In the following, $\theta_{j,k,l}$ will denote the value v_l of $g_{j,k}$. Together with the transition scores ϕ, the values at the supporting points $\theta_{j,k,l}$ constituted the parametrization of the model (in the following collectively denoted by $\boldsymbol{\theta}$).

Learning algorithm

Let n be the number of training examples $(\mathbf{x}^{(i)}, \boldsymbol{\pi}^{(i)})$, $i = 1,\ldots,n$. Following the discriminative learning paradigm, we wanted to enforce a large margin of separation between the correct path $\boldsymbol{\pi}^{(i)}$ and *any* other wrong path $\overline{\boldsymbol{\pi}} \neq \boldsymbol{\pi}^{(i)}$, i.e.,

$$F_{\boldsymbol{\theta}}(\mathbf{x}^{(i)},\boldsymbol{\pi}^{(i)}) - F_{\boldsymbol{\theta}}(\mathbf{x}^{(i)},\overline{\boldsymbol{\pi}}) \gg 0 \quad \forall \overline{\boldsymbol{\pi}} \neq \boldsymbol{\pi}^{(i)} \quad \forall i = 1,\ldots,n.$$

To achieve this, the following linear programming problem (LP) is solved:

$$\min_{\boldsymbol{\theta},\boldsymbol{\xi}\ge\mathbf{0}} \frac{1}{n}\sum_{i=1}^{n}\xi^{(i)} + C\Omega(\boldsymbol{\theta})$$

s.t.

$$F_{\boldsymbol{\theta}}(\mathbf{x}^{(i)},\boldsymbol{\pi}^{(i)}) - F_{\boldsymbol{\theta}}(\mathbf{x}^{(i)},\overline{\boldsymbol{\pi}}) \ge 1 - \xi^{(i)} \quad \forall \overline{\boldsymbol{\pi}} \neq \boldsymbol{\pi}^{(i)} \quad \forall i = 1,\ldots,n, \qquad (1)$$

where Ω is a linear regularization term of the form

$$\Omega(\boldsymbol{\theta}) = |\boldsymbol{\theta}| + \sum_{j=1}^{m}\sum_{k\in S}\sum_{l=1}^{Q-1}|\theta_{j,k,l} - \theta_{j,k,l+1}|.$$

Note that $F_{\boldsymbol{\theta}}$ is linear in all parameters and hence the constraints in Equation 1 are linear. Regularization is a technique commonly used in empirical inference to avoid overfitting. Our regularizer implements the idea that absolute parameter values should be small, and it penalizes the variation of the feature scoring functions (with respect to the choice of supporting points). Regularization strength can be adjusted using the hyper-parameter C.

We introduced so-called slack variables $\xi^{(i)}$ to implement a soft-margin (Cortes and Vapnik 1995) allowing some prediction errors on the training set. As there are exponentially many wrong paths $\overline{\boldsymbol{\pi}}$, we also have an exponential number of margin constraints in Equation 1. This prohibits solving the optimization problem directly. Instead, starting from an empty set of margin constraints and a random parametrization $\boldsymbol{\theta}^{(1)}$, for every training example we computed the wrong path that maximally violates the margin constraints. We used a generalized Viterbi algorithm that decodes the two best paths, thereby allowing us to identify the wrong path (since there is only one correct path). Adopting a column generation technique, adding constraints and solving the intermediate LP were iterated till convergence to the (provably) optimal solution (Hettich and Kortanek 1993; Rätsch et al. 2002; Altun et al. 2003): At iteration t, new margin constraints of the form

$$F_{\boldsymbol{\theta}^{(t)}}(\mathbf{x}^{(i)},\boldsymbol{\pi}^{(i)}) - \max_{\overline{\boldsymbol{\pi}}\neq\boldsymbol{\pi}^{(i)}}\{F_{\boldsymbol{\theta}^{(t)}}(\mathbf{x}^{(i)},\overline{\boldsymbol{\pi}})\} \ge 1 - \xi^{(i)}$$

were added to the problem, which was solved again to obtain the next intermediate solution $\boldsymbol{\theta}^{(t+1)}$. The intermediate LPs were solved using the CPLEX optimization software (http://www.ilog.com/products/cplex/), which facilitated training with $n = 12{,}000$ examples.

Loss function

We augmented the basic algorithm described above with a loss function Δ that adjusts the loss a path incurs depending on its similarity to the true path. That is, a path that closely resembles the truth incurs a small loss compared to one that is completely different from the true path. The loss function is used to rescale the margin (Altun et al. 2003; Taskar et al. 2003), replacing the margin constraints in Equation 1 with

$$F_{\boldsymbol{\theta}}(\mathbf{x}^{(i)},\boldsymbol{\pi}^{(i)}) - F_{\boldsymbol{\theta}}(\mathbf{x}^{(i)},\overline{\boldsymbol{\pi}}) \ge \Delta(\boldsymbol{\pi}^{(i)},\overline{\boldsymbol{\pi}}) - \xi^{(i)} \quad \forall \overline{\boldsymbol{\pi}} \neq \boldsymbol{\pi}^{(i)}, i = 1,\ldots,n.$$

During optimization, the loss was taken into account when decoding to find the maximal margin violator:

$$\operatorname*{argmax}_{\overline{\boldsymbol{\pi}}\neq\boldsymbol{\pi}^{(i)}}\{F_{\boldsymbol{\theta}}(\mathbf{x}^{(i)},\overline{\boldsymbol{\pi}}) + \Delta(\boldsymbol{\pi}^{(i)},\overline{\boldsymbol{\pi}})\}.$$

The loss was required to be non-negative and decomposable for efficient decoding via dynamic programming. We chose a position-wise loss $\ell(p)$, which is summed over the whole sequence (of length t): $\Delta = \sum_{p=1}^{t}\ell(p)$ (similar to a weighted Hamming loss; for details, cf. Supplemental Table S7).

Cross-validation, evaluation, and whole-genome predictions

For fivefold cross-validation, fragments in the 2010 set were randomly split into five subsets, where we ensured that across all accessions overlapping sequences were assigned to the same subset. The first predictor was trained on the first three subsets, its optimal regularization parameter C was selected on the fourth subset, and its performance was evaluated on the fifth subset. For the other four predictors, the assignment of training, validation, and test set was permuted in order to obtain unbiased (test) predictions for all 2010 data.

All evaluations were based on data from 18 accessions (no predictions were made for the reference, and for Van-0 no reliably labeled set exists; Clark et al. 2007). Furthermore, known PRs as well as predicted PRs were excluded from specificity–sensitivity estimation if they contained ≥75% repetitive sites.

Replacing transition scores $\phi(i,i)$, $i \in \{C_U,C_R\}$ after training by $\hat{\phi}(i,i) = \phi(i,i) + \delta$ resulted in predictions either with increased specificity ($\delta > 0$) or with increased sensitivity ($\delta < 0$). Fifty-one values for δ were uniformly chosen from the interval $[-3, 2]$ to generate specificity-sensitivity curves for all five test subsets. For Figure 2 and Supplemental Figure S3, specificity-sensitivity curves were averaged over the subsets.

The sequence type of each nucleotide was determined based

on the TAIR6 *A. thaliana* genome annotation available at http://www.arabidopsis.org. In cases where annotations overlapped, the sequence type was assigned following the hierarchy: coding > UTR/intron > intergenic. PRs were assigned a sequence type based on the majority of nucleotides contained.

Specificity and sensitivity for whole-genome predictions are expected to be slightly different from the values estimated on the 2010 set as coding sequences are relatively overrepresented for 2010 compared with the entire genome (Nordborg et al. 2005; Clark et al. 2007). To account for the compositional bias of the 2010 data, we applied the following correction: Let n^T_{cod} be the number of coding bases in the 2010 data; n^G_{cod}, the number of coding bases in the genome. Then, for the whole genome the number of TPs in coding regions is estimated as $TP^G_{cod} = (n^G_{cod}/n^{MN}_{cod})TP^T_{cod}$. Applying the same corrections for *FP*s, *TD*s, and *FN*s, as well as for intergenic (*ige*) and UTR/intron bases (*utr*), specificity was recalculated as

$$\frac{TP^G_{cod} + TP^G_{ige} + TP^G_{utr}}{TP^G_{cod} + TP^G_{ige} + TP^G_{utr} + FP^G_{cod} + FP^G_{ige} + FP^G_{utr}}$$

and sensitivity as

$$\frac{TD^G_{cod} + TD^G_{ige} + TD^G_{utr}}{TD^G_{cod} + TD^G_{ige} + TD^G_{utr} + FN^G_{cod} + FN^G_{ige} + FN^G_{utr}}.$$

To obtain PR predictions with high specificity, transition scores were independently tuned by choosing the smallest δ for which each of the predictors achieved specificity ≥90% on its test set. Whole-genome predictions were made independently with every predictor, and a single prediction was assigned to every position according to the following scheme: The genome was partitioned into chunks of ~1 kb (breakpoints between chunks were only set where all five predictors agreed on C_U or C_R). If a chunk contained a 2010 sequence fragment, the respective test predictions were used. Otherwise one of the five predictors was chosen randomly for the given chunk.

Methods for the analyses of genome-wide predictions, experimental characterization of predictions, evaluation of genome-wide patterns of polymorphism, overlap of predictions relative to *cis*-regulatory sites, and annotation of predictions relative to genes are provided in the Supplemental material. GenBank accession numbers for *RPM1* are ET181618–ET181629. The software used to produce the results in the article and an open source toolbox with an improved and easier-to-use implementation of the algorithm is available at http://www.fml.tuebingen.mpg.de/raetsch/projects/mppr.

Acknowledgments

We thank Gabriele Schweikert, Bernhard Schölkopf, and Stephan Ossowski for helpful discussions and Stephan Ossowski for assistance in visualizing predictions. Supported by Innovation Funds and core funding of the Max Planck Society. D.W. is a director of the Max Planck Institute. G.R. is a group leader at the Friedrich Miescher Laboratory of the Max Planck Society.

References

Alonso, J. and Ecker, J. 2006. Moving forward in reverse: Genetic technologies to enable genome-wide phenomic screens in *Arabidopsis*. *Nat. Rev. Genet.* **7:** 524–536.

Altun, Y., Tsochantaridis, I., and Hofmann, T. 2003. Hidden Markov support vector machines. In *Proceedings of the 20th International Conference on Machine Learning*, pp. 3–10. AAAI Press, Menlo Park, CA.

The *Arabidopsis* Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408:** 796–815.

Bakker, E., Toomajian, C., Kreitman, M., and Bergelson, J. 2006. A genome-wide survey of R gene polymorphisms in *Arabidopsis*. *Plant Cell* **18:** 1803–1818.

Bernal, A., Crammer, K., Hatzigeorgiou, A., and Pereira, F. 2007. Global discriminative learning for higher-accuracy computational gene prediction. *PLoS Comput. Biol.* **3:** e54. doi: 10.1371/journal.pcbi.0030054.

Bolstad, B., Irizarry, R., Astrand, M., and Speed, T. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19:** 185–193.

Borevitz, J., Hazen, S., Michael, T., Morris, G., Baxter, I., Hu, T., Chen, H., Werner, J., Nordborg, M., Salt, D., et al. 2007. Genome-wide patterns of single-feature polymorphism in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci.* **104:** 12057–12062.

Burge, C. and Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268:** 78–94.

Chee, M., Yang, R., Hubbell, E., Berno, A., Huang, X., Stern, D., Winkler, J., Lockhart, D., Morris, M., Fodor, S., et al. 1996. Accessing genetic information with high-density DNA arrays. *Science* **274:** 610–614.

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

Cortes, C. and Vapnik, V. 1995. Support vector networks. *Mach. Learn.* **20:** 273–297.

Cutler, D., Zwick, M., Carrasquillo, M., Yohn, C., Tobin, K., Kashuk, C., Mathews, D., Shah, N., Eichler, E.J.W., Warrington, J.A., et al. 2001. High-throughput variation detection and genotyping using microarrays. *Genome Res.* **11:** 1913–1925.

Dawson, E., Chen, Y., Hunt, S., Smink, L., Hunt, A., Rice, K., Livingston, S., Bumpstead, S., Bruskiewich, R., Sham, P., et al. 2001. A SNP resource for human chromosome 22: Extracting dense clusters of SNPs from the genomic sequence. *Genome Res.* **11:** 170–178.

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. 1998. *Biological sequence analysis*: *Probabilistic models of protein and nucleic acids*, 7th ed. Cambridge University Press, Cambridge.

Fahlgren, N., Howell, M., Kasschau, K., Chapman, E., Sullivan, C., Cumbie, J., Givan, S., Law, T., Grant, S., Dangl, J., et al. 2007. Highthroughput sequencing of Arabidopsis microRNAs: Evidence for frequent birth and death of miRNA genes. *PLoS ONE* **1:** e14. doi: 10.1371/journal.pone.0000219.

Frazer, K., Wade, C., Hinds, D., Patil, N., Cox, D., and Daly, M. 2004. Segmental phylogenetic relationships of inbred mouse strains revealed by finescale analysis of sequence variation across 4.6 Mb of mouse genome. *Genome Res.* **14:** 1493–1500.

Frazer, K., Eleazar, E., Kang, H., Bogue, M., Hinds, D., Beilharz, E., Gupta, R., Montgomery, J., Morenzoni, M., Nilsen, G., et al. 2007. A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. *Nature* **448:** 1050–1053.

Giegerich, R., Meyer, C., and Steffen, P. 2004. A discipline of dynamic programming over sequence data. *Sci. Comput. Program.* **51:** 215–263.

Grant, M., Godiard, L., Straube, E., Ashfield, T., Lewald, J., Sattler, A., Innes, R., and Dangl, J. 1995. Structure of the *Arabidopsis* RPM1 gene enabling dual specificity disease resistance. *Science* **269:** 843–846.

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*. *Proc. Natl. Acad. Sci.* **95:** 15843–15848.

Hettich, R. and Kortanek, K. 1993. Semi-infinite programming: Theory, methods and applications. *SIAM Rev.* **3:** 380–429.

Hinds, D., Stuve, L., Nilsen, G., Halperin, E., Eskin, E., Ballinger, D., Frazer, K., and Cox, D. 2005. Whole-genome patterns of common DNA variation in three human populations. *Science* **307:** 1072–1079.

Hinds, D., Kloek, A., Jen, M., Chen, X., and Frazer, K. 2006. Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nat. Genet.* **38:** 82–85.

The International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* **437:** 1299–1320.

Johanson, U., West, J., Lister, C., Michaels, S., Amasino, R., and Dean, C. 2000. Molecular analysis of FRIGIDA, a major determinant of natural variation in *Arabidopsis* flowering time. *Science* **290:** 344–347.

Jones, J. and Dangl, J. 2006. The plant immune system. *Nature* **444:** 323–329.

Jones-Rhoades, M., Bartel, D., and Bartel, B. 2006. MicroRNAs and their regulatory roles in plants. *Annu. Rev. Plant Biol.* **57:** 19–53.

Kim, S., Plagnol, V., Hu, T., Toomajian, C., Clark, R., Ossowski, S., Ecker, J., Weigel, D., and Nordborg, M. 2007. Recombination and linkage disequilibrium in *Arabidopsis thaliana*. *Nat. Genet.* **39:** 1151–1155.

Lafferty, J., McCallum, A., and Pereira, F. 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA.

Lee, I., Dombkowski, A., and Athey, B. 2004. Guidelines for incorporating non-perfectly matched oligonucleotides into target-specific hybridization probes for a DNA microarray. *Nucleic Acids Res.* **32:** 681–690.

Lee, J.-Y., Colinas, J., Wang, J., Mace, D., Ohler, U., and Benfey, P. 2006. Transcriptional and posttranscriptional regulation of transcription factor expression in Arabidopsis roots. *Proc. Natl. Acad. Sci.* **103:** 6055–6060.

McNally, K., Bruskiewich, R., Mackill, D., Buell, C., Leach, J., and Leung, H. 2006. Sequencing multiple and diverse rice varieties. connecting whole genome variation with phenotypes. *Plant Physiol.* **141:** 26–31.

Mills, R., Luttig, C., Larkins, C., Beauchamp, A., Tsui, C., Pittard, W., and Devine, S. 2006. An initial map of insertion and deletion (indel) variation in the human genome. *Genome Res.* **16:** 1182–1190.

Müller, K.-R., Mika, S., Rätsch, G., Tsuda, K., and Schölkopf, B. 2001. An introduction to kernel-based learning algorithms. *IEEE Trans. Neural Netw.* **12:** 181–201.

Mural, R., Adams, M., Myers, E., Smith, H., Miklos, G., Wides, R., Halpern, A., Li, P., Sutton, G., Nadeau, J.A., et al. 2002. A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. *Science* **296:** 1661–1671.

Nguyen, N. and Guo, Y. 2007. Comparisons of sequence labeling algorithms and extensions. In *Proceedings of the 24th International Conference on Machine Learning*, pp. 681–688. ACM Press, Corvalis, OR.

Nordborg, M., Hu, T., Ishino, Y., Jhaveri, J., Toomajian, C., Zheng, H., Bakker, E., Calabrese, P., Gladstone, J., Goyal, R., et al., 2005. The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biol.* **3:** e196. doi: 10.1371/journal.pbio.0030196.

O'Connor, T., Dyreson, C., and Wyrick, J. 2005. Athena: A resource for rapid visualization and systematic analysis of *Arabidopsis* promoter sequences. *Bioinformatics* **21:** 4411–4413.

Patil, N., Berno, A., Hinds, D., Barrett, W., Doshi, J., Hacker, C., Kautzer, C., Lee, D., Marjoribanks, C., McDonough, D., et al. 2001. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294:** 1719–1723.

Rajagopalan, R., Vaucheret, H., Trejo, J., and Bartel, D. 2006. A diverse and evolutionarily fluid set of microRNAs in *Arabidopsis thaliana*. *Genes & Dev.* **20:** 3407–3425.

Rätsch, G. and Sonnenburg, S. 2007. Large scale hidden semi-markov SVMs. In *Advances in neural information processing systems 19* (eds. B. Schölkopf et al.), pp. 1161–1168. MIT Press, Cambridge, MA.

Rätsch, G., Demiriz, A., and Bennett, K. 2002. Sparse regression ensembles in infinite and finite hypothesis spaces. *Mach. Learn.* **48:** 193–221.

Rätsch, G., Sonnenburg, S., Srinivasan, J., Witte, H., Müller, K.-R., Sommer, R., and Schölkopf, B. 2007. Improving the *Caenorhabditis elegans* genome annotation using machine learning. *PLoS Comput. Biol.* **3:** e20. doi: 10.1371/journal.pcbi.0030020.

Schmid, K., Sorensen, T., Stracke, R., Torjek, O., Altmann, T., Mitchell-Olds, T., and Weisshaar, B. 2003. Large-scale identification and analysis of genome-wide single-nucleotide polymorphisms for mapping in *Arabidopsis thaliana*. *Genome Res.* **13:** 1250–1257.

Schmid, K., Ramos-Onsins, S., Ringys-Beckstein, H., Weisshaar, B., and Mitchell-Olds, T. 2005. A multilocus sequence survey in *Arabidopsis thaliana* reveals a genome-wide departure from a neutral model of DNA sequence polymorphism. *Genetics* **169:** 1601–1615.

Schölkopf, B. and Smola, A. 2002. *Learning with kernels*. MIT Press, Cambridge, MA.

Schulze, U., Hepp, B., Ong, C., and Rätsch, G. 2007. PALMA: mRNA to genome alignments using large margin algorithms. *Bioinformatics*. **23:** 1892–1890.

Sha, F. and Saul, L. 2007. Large margin hidden Markov models for automatic speech recognition. In *Advances in neural information processing systems 19*, (eds. B. Schölkopf et al.), pp. 1249–1256. MIT Press, Cambridge, MA.

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:** 1243–1250.

Shendure, J., Mitra, R., Varma, C., and Church, G. 2004. Advanced sequencing technologies: Methods and goals. *Nat. Rev. Genet.* **5:** 335–344.

Song, W.-Y., Wang, G.-L., Chen, L.-L., Kim, H.-S., Pi, L.-Y., Holsten, T., Gardner, J., Wang, B., Zhai, W.-X., Zhu, L.-H., et al. 1995. A receptor kinase-like protein encoded by the rice disease resistance gene, Xa21. *Science* **270:** 1804–1806.

Taskar, B., Guestrin, C., and Koller, D. 2004. Max-margin Markov networks. In *Advances in neural information processing systems 16* (eds. S. Thrun et al.), pp. 25–32. MIT Press, Cambridge, MA.

Thomas, J. 2006. Adaptive evolution in two large families of ubiquitin ligase adapters in nematodes and plants. *Genome Res.* **16:** 1017–1030.

Toomajian, C., Hu, T., Aranzana, M., Lister, C., Tang, C., Zheng, H., Zhao, K., Calabrese, P., Dean, C., and Nordborg, M., et al. 2006. A nonparametric test reveals selection for rapid flowering in the *Arabidopsis* genome. *PLoS Biol.* **4:** e137. doi: 10.1371/journal.pbio.0040137.

Tsochantaridis, I., Joachims, T., Hofmann, T., and Altun, Y. 2005. Large margin methods for structured and interdependent output variables. *J. Mach. Learn. Res.* **6:** 1453–1484.

Tuskan, G., DiFazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A., et al. 2006. The genome of black cottonwood, *Populus trichocarpa*. *Science* **313:** 1596–1604.

Vapnik, V. 1995. *The nature of statistical learning theory*. Springer Verlag, New York.

Wicks, S., Yeh, R., Gish, W., Waterston, R., and Plasterk, R. 2001. Rapid gene mapping in *Caenorhabditis elegans* using a high density polymorphism map. *Nat. Genet.* **28:** 160–164.

Windsor, A., Schranz, M., Formanova, N., Gebauer-Jung, S., Bishop, J., Schnabelrauch, D., Kroymann, J., and Mitchell-Olds, T. 2006. Partial shotgun sequencing of the *Boechera stricta* genome reveals extensive microsynteny and promoter conservation with *Arabidopsis*. *Plant Physiol.* **140:** 1169–1182.

Wright, S. and Gaut, B. 2005. Molecular population genetics and the search for adaptive evolution in plants. *Mol. Biol. Evol.* **22:** 506–519.

Received August 9, 2007; accepted in revised form March 5, 2008.