Machine Learning Algorithms for the Analysis of Data from Whole-Genome Tiling Microarrays

Dissertation
der Fakultät für Informations- und Kognitionswissenschaften
der Eberhard-Karls-Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Dipl.-Inform. (Bioinf.) Georg F. Zeller
aus Konstanz

Tübingen
2009

Tag der mündlichen Qualifikation:	21.04.2010
Dekan:	Prof. Dr.-Ing. Oliver Kohlbacher
1. Berichterstatter:	Prof. Dr. Daniel H. Huson
2. Berichterstatter:	Prof. Dr. Detlef Weigel
3. Berichterstatter:	Prof. Dr. Klaus-Robert Müller

To my father

Erklärung

Hiermit erkläre ich, dass ich diese Schrift selbständig und nur mit den angegebenen Hilfsmitteln angefertigt habe und dass alle Stellen, die im Wortlaut oder dem Sinne nach anderen Werken entnommen sind, durch Angaben der Quellen kenntlich gemacht sind.

Tübingen, Oktober 2009 *Georg Zeller*

Abstract

In this work we developed machine learning-based methods with the aim to further our understanding regarding fundamental questions of molecular biology, using as our example the model plant *Arabidopsis thaliana*:

What are the differences between genomes of individuals belonging to the same species? Characterizing sequence variants (polymorphisms) genome-wide is a prerequisite for establishing causal links between adaptive quantitative traits and the underlying genetic variants. Single-nucleotide polymorphisms (SNPs) are the most abundant class of polymorphisms. In addition to SNP detection, we investigated genomic regions in which SNP calling algorithms tend to fail: on the one hand, highly variable sequence tracts, for which, paradoxically, only very few SNPs can be identified and, on the other hand, additional polymorphism types, such as insertions and deletions. With our newly developed method (mPPR) we discovered hundreds of thousands of polymorphic regions (with a false-discovery rate of $< 3\%$). These correspond, in part, to SNPs, but also contain deletions ranging from a few to several thousand nucleotides in length. Our results revealed, for the first time, a comprehensive, fine-scale picture of the polymorphism patterns in *A. thaliana* with dramatic differences between coding and noncoding regions and also between individual genes and gene families.

What is an organism's full complement of genes, in which tissues and developmental stages are they transcribed and how is their expression altered in response to environmental changes? Transcriptome studies have provided the foundation for reconstruction of the gene regulatory network, which describes the control of cellular processes, e.g., during cell differentiation. We developed a transcript identification method (mSTAD), which recognizes genic expression patterns. With mSTAD, we discovered thousands of new transcripts that were not previously known despite extensive annotation efforts. Validation experiments confirmed $> 75\%$ of the tested cases, corroborating mSTAD's high accuracy. Moreover, we found hundreds of genomic regions with evidence of stress-specific transcription. These include previously unannotated genes as well as wrongly annotated parts of known genes.

Our computational methods are based on data generated with so-called tiling arrays, an advanced DNA microarray which interrogates a whole genome in regular intervals. It facilitates both the detection of polymorphisms and transcriptome profiling. Using this technology our analyses targeted, for the first time, the whole genome and were not restricted to a few fragments.

Since the resulting data resources are the basis for further research, high accuracy was imperative. However, microarray data typically exhibits high noise levels. We therefore devised new preprocessing techniques to reduce systematic noise, in particular probe sequence effects. We demonstrated the benefit of this technique for subsequent transcript identification. In contrast to that, comparable methods investigated here failed in this aspect. In our attempts to detect polymorphic or transcribed regions, we were facing segmentation problems. Recently developed machine learning algorithms, especially Hidden Markov Support Vector Machines, were found to be very well-suited for solving these problems. In the case of transcript identification, we could show mSTAD's superior accuracy compared to other widely used methods. Since no comparable methods exist for polymorphic region prediction, however, no such comparison was possible. Although originally developed for the analysis of *A. thaliana* data, our methods can nevertheless be broadly applied to similar data sets, which already exist for a number of organisms. We furthermore discuss their applicability to related data as it is, for instance, being generated by next-generation sequencing technologies.

Keywords

abiotic stress, array-based resequencing, Arabidopsis, expression analysis, genome annotation, hidden Markov model, hidden Markov support vector machine, machine learning, natural variation, polymorphic region, polymorphism discovery, tiling array, transcriptome, transcript identification, transfrag

Zusammenfassung

Im Rahmen dieser Dissertation wurden auf maschinellen Lerntechniken basierende, bioinformatische Methoden entwickelt, um den Kenntnisstand in Bezug auf zentrale molekularbiologische Fragen am Beispiel der Modellpflanze *Arabidopsis thaliana* zu erweitern:

Inwiefern unterscheiden sich die Genome einzelner Individuen derselben Spezies? *Sequenzvariation* (Polymorphismen) im großen Stil zu charakterisieren ist die Voraussetzung, um adaptive, quantitative phänotypische Merkmale auf die ursächlichen genetischen Varianten zurückführen zu können. Die häufigste Klasse von Sequenzvarianten sind Einzelnukleotidänderungen (SNPs). Neben der Erkennung von SNPs untersuchten wir Genombereiche genauer, in denen SNP-Erkennungsverfahren nur unzureichend funtionieren: Einerseits hochvariable Regionen, für die paradoxerweise nur sehr wenige SNPs identifiziert werden können, und andererseits weitere Varianten, wie Insertionen und Deletionen. Mit unserer neu entwickelten Methode (mPPR) fanden wir hunderttausende *polymorphe Regionen* (unter denen wir $< 3\%$ Falschpositive erwarten), die teils SNPs beinhalten, teils Deletionen mit einigen wenigen bis zu tausenden von Nukleotiden. Aus diesen Resultaten entstand erstmal ein umfassendes, hochaufgelöstes Bild der Polymorphismenmuster in *Arabidopsis*, mit drastischen Unterschieden zwischen kodierenden und nichtkodierenden Bereichen, aber auch zwischen einzelnen Genen und Genfamilien.

Wie sieht die Gesamtheit der Gene eines Organismus' aus, in welchen Geweben und Entwicklungsstadien werden sie transkribiert, und wie verändert sich ihre Expression unter Umwelteinflüssen? Entsprechende *Transkriptomanalysen* bilden die Basis zur Rekonstruktion des Genregulationsnetzwerks, welches die Steuerung zellulärer Prozesse, z.B. der Zelldifferenzierung, beschreibt. Wir entwickelten ein Verfahren zur *Transkriptsuche* (mSTAD), das Gene aufgrund von Expressionsmessungen erkennen kann. Damit identifizierten wir tausende neue Transkripte, die ungeachtet großer vorhergehender Annotationsprojekte bisher unbekannt waren. Durch Validierungsexperimente konnten $> 75\%$ der Kandidaten bestätigt und so mSTAD's Genauigkeit experimentell belegt werden. Darüber hinaus fanden wir hunderte von genomischen Regionen, die spezifisch unter Stressbedingungen transkribiert werden. Sie umfassen sowohl zuvor unbekannte Gene, als auch bisher fehlerhaft annotierte Bereiche bereits bekannter Gene.

Unsere bioinformatischen Methoden basieren auf Daten von sogenannten *Tiling-Arrays*, einer hochentwickelten DNS-Microarray-Technologie, die durch genomweite Messungen in einem feinen Raster die Detektion von Genomvariation sowie Transkriptomanalysen ermöglicht. So konnten wir erstmals das ganze Genom untersuchen und mussten uns nicht auf wenige Fragmente beschränken.

Da unsere Resultate die Grundlage für weitergehende Forschung bilden, ist *hohe Genauigkeit* der Analysen von größter Bedeutung. Microarray-Daten kennzeichnet jedoch typischerweise starkes Rauschen. Wir entwickelten deshalb neue Vorverarbeitungstechniken um systematisches Rauschen, insbesondere Sondensequenzeffekte, zu verringern. Wir zeigten den klaren Nutzen dieser Technik für anschließende Transkripterkennung. Vergleichbare, hier untersuchte Vorverarbeitungsmethoden versagten hingegen unter diesem zentralen Gesichtspunkt. Bei der Erkennung polymorpher Regionen oder transkribierter Bereiche sind wir mit *Segmentationspoblemen* konfrontiert, die sich mit kürzlich entwickelten maschinellen Lernmethoden, insbesondere den *Hidden Markov Support Vector Machines*, sehr gut lösen lassen. Im Falle der Transkriptsuche konnten wir mSTAD's überlegene Genauigkeit im Vergleich zu anderen gängigen Analysetechniken empirisch belegen, wohingegen zur Erkennung polymorpher Regionen keine konkurrierenden Methoden existierten. Obwohl für *Arabidopsis*-Daten entwickelt, sind unsere Methoden anwendbar auf vergleichbare Datensätze, die für viele weitere Organismen existieren. Wir diskutieren ferner ihre Eignung für die Analyse verwandter Daten, wie sie z.B. mit neuen Sequenzierungstechniken erzeugt werden.

Acknowledgements

To my advisors Gunnar Rätsch and Detlef Weigel I am very thankful – not only for many fruitful discussions, ideas gratefully adopted for this work, and general advice, but also for creating an excellent research environment. I truly enjoyed working in a friendly and open atmosphere on the Max Planck Campus in Tübingen.

Additionally, I would like to express my thanks to the members of my thesis committee, Daniel Huson, to whom I am also very grateful for his long-term support since undergraduate times, Detlef Weigel, and Klaus-Robert Müller for providing invaluable stimuli from the perspective of a scientist working on a – seemingly – very different interface between machine learning and biology.

I am extremely thankful to Richard M. Clark and Sascha Laubinger for working with me in a very open-minded and productive manner on a number of projects. It has been great fun, and their contributions to the work on which this thesis is based were distinctive.

Moreover, I would like to thank colleagues who provided (unpublished) data or source code for this thesis: Gunnar Rätsch, Jonas Behr, Regina Bohnert, Jun Cao, Cheng Soon Ong, Stephan Ossowski, Korbinian Schneeberger, and Fabio De Bona.

Further, I would like to thank my colleagues, collaborators and coauthors. Working together has been a pleasure and honor for me, and without their contributions, neither this thesis nor any of my publications would exist: Sascha Laubinger, Richard M. Clark, Gabriele Schweikert, Stefan Henz, Stephan Ossowski, Korbinian Schneeberger, Regina Bohnert, Jonas Behr, Christian Widmer, Alexander Zien, Sören Sonnenburg, Timo Sachsenberg, Wolfgang Busch, Fabio De Bona, Cheng Soon Ong, Petra Philips, Norman Warthmann, Anja Bohlen, Lisa Hartmann, Nina Krüger, Naira Naouar, Tina T. Hu, Kevin L. Childs, and many more. In particular, I benefited from discussions with and insights shared by Gunnar Rätsch, Detlef Weigel, Bernhard Schölkopf, Jan Lohmann, and Magnus Nordborg.

I am furthermore thankful to Gunnar Rätsch, Ulrike Winter, Jonas Behr, Sascha Laubinger, Detlef Weigel, and Marc Bégin for critically reading the thesis manuscript. Their suggestions and corrections helped to improve it substantially.

Thanks go to Regina Bohnert and Johannes Eichner. I had fun and learned a lot by co-advising their *Diplom* projects.

Also, I would like to acknowledge the people in Gunnar's group and in Detlef's lab for many inspiring discussions (not only during "bio-breakfasts" etc.). The intellectual environment formed by all of them gave me a taste of what it could mean to be a scientist.

The computational experiments conducted for this thesis profited a lot by an exceptional computing environment on the Max Planck Campus. My special thanks go to our administrators Andre Noll and Sebastian Stark for installing and maintaining it.

I very much enjoyed scientific discussions in the broadest sense that I had with many people (including all of the aforementioned) during my time as a PhD student. Here, I would like to add my thanks to Johannes Söding, Timothy Davison, Tobias Klöpper, and Nickias Kienle.

Additionally, I would like to thank my fellow students in Tübingen and Uppsala, my

teachers at these Universities and many people I met at research conferences; not only for what I have learned during talks and discussions, but more importantly for fostering my general enthusiasm about computational biology.

Moreover, I gratefully acknowledge funding from the Max Planck Society and the SIROCCO EU Integrated Project.

Last but not least, I would like to express my deep gratitude to my parents and my family for their constant support.

Contents

1 Introduction

The publication of a human reference genome sequence [95] has fundamentally transformed research in the life sciences. Notwithstanding that this was a milestone scientific achievement and truly a success for bioinformatics, it marked the beginning rather than the end of genome sequencing efforts. The idea that every human being is a unique individual is, in part, reflected by differences between our individual genome sequences. It has thus become increasingly clear that without information on DNA variants, which distinguish individuals and subpopulations, our knowledge of the human genome is incomplete. In other words, the problem of deciphering the genome has been reframed as characterizing *DNA sequence variation*. Discovering polymorphisms, i.e., sites where individual genomes differ, has thus been a prime objective in the life sciences [162].

Although all cells of an organism share the same genome, i.e., contain the same hereditary information, their morphology and physiology exhibit a tremendous diversity. How this diversity across different organs and developmental stages is reconstituted in a tightly regulated manner in each individual has largely remained a mystery for complex organisms, although it has been an absolutely central research question. As the first step, we seek to understand which parts of the genome constitute genes that are expressed, i.e., transcribed into RNA. These transcripts can in turn perform cellular functions or be further translated into proteins. Because gene expression is a highly dynamic process, we moreover aim at monitoring when during development, where in the organism and at which rate, certain genes are expressed. Eventually, comprehensive gene expression data will facilitate to disentangle the regulatory network underlying the observed expression patterns. An organism's full complement of transcripts is also referred to as the transcriptome, and the term *transcriptomics* was coined for the research area revolving around these problems.

Many experimental techniques have been developed to address these fundamental genomics and transcriptomics challenges. When I started my graduate studies, the most advanced hybridization-based technique, *tiling microarrays*, was the leading technology for comprehensive and relatively cost-effective whole-genome measurements, with diverse applicability, including expression profiling and polymorphism discovery (or array-based resequencing) [117]. However, despite extensive research analyzing microarray data, which is characterized by relatively high noise levels, many challenges have remained. In this context, methods from statistics and *machine learning* have proven particularly well-suited; for instance, the first application of support vector machines (SVMs) within the field of computational biology was an analysis of microarray data [18].

In the following section, we will continue with a brief overview of some machine learning concepts relevant to our work. Subsequently, tiling microarrays will be introduced in detail, including a review of previous work on array-based transcriptomics and resequencing. Furthermore, we will discuss alternative experimental approaches and outline which computational methods we have developed for analyzing tiling arrays. In the final introductory sections, we will present the model plant *Arabidopsis thaliana* and highlight how our results have2 contributed to the understanding of its genome and transcriptome.

1.1 A Machine Learning Primer

In essence, machine learning algorithms are statistical models able to make accurate predictions (usually denoted by y) from example data (denoted by x). We talk about *learning* or *inference*, if such a model is adaptive, i.e., has free parameters that are estimated from the data in a process called *training* and is then able to make predictions on examples it has not previously seen. Based on our understanding of the problem at hand, we sometimes extract relevant information in the form of *features* as input for the learning algorithm instead of directly using primary data. Depending on the nature of the predictions, three problems commonly encountered in machine learning (and also within the scope of this thesis) can be identified:

- If interested in classifying data into two classes, i.e., $y \in \{-1, +1\}$, one uses algorithms for *binary classification.* The classification scenario can also be generalized to the case of multiple classes ($y \in \{1, 2, \ldots, k\}$ for some finite $k \in \mathbb{N}$).
- If the prediction target is a continuous variable, i.e., $y \in \mathbb{R}$, the learning task is called *regression.*
- When dealing with sequential data, where each example is itself a sequence of data points ($x = x_1, \ldots, x_n$ of length $n \in \mathbb{N}$), one is often interested in predicting a *label sequence* ($y = y_1, \ldots, y_n$) for each example that assigns a (scalar) class label to each position in x. If additional restrictions on valid label sequences exist or if one wants to exploit global properties of the example sequences, label sequence learning problems cannot simply be reformulated as multi-class classification problems.

The key contributions of this thesis are based on *supervised* machine learning algorithms. In the supervised learning scenario, the algorithm is provided with a so-called set of *labeled data*, consisting of examples x together with corresponding labels y indicating what the correct predictions should be. In genome research, we often encounter the following situation: Whole-genome measurements are available from high-throughput assays, but require additional automatic interpretation to extract biologically relevant information, e.g., to identify SNPs or transcripts. Additionally, a few genes or regions have been characterized in detail by in-depth experiments, which allow us to extract label information. For instance, in its pilot phase, the ENCODE project focused experimental resources on regions covering 1% of the human genome to create high-quality annotations [30]. Our main motivation for the application of machine learning algorithms is to exploit this labeled data for training predictive models that are subsequently able to make accurate predictions for the whole genome.

During training, a supervised learning algorithm is provided with data-derived input features and associated labels allowing it to recognize properties that are predictive of a certain label. Once trained, it should be able to predict the label when it is only given the features. Only if its predictions on previously unseen examples are accurate, does it exhibit a good *generalization* performance; otherwise, if the model only fits the training data well, it is said to *overfit*. Hence, to assess the predictive power of a learning algorithm, one also needs a set of labeled examples that is disjoint from the training set. It is common practice to evaluate the performance of a learning algorithm by means of *cross-validation*. For this, the labeled data set is partitioned into subsets that are either used for training or for validation, i.e., performance assessment. If hyperparameters of the learning algorithm have to be adjusted, a third set can be reserved for *model selection*. The assignment of subsets into training and test sets is then permuted until all data points have been used once for validation; the validation error is reported by averaging across permutations.

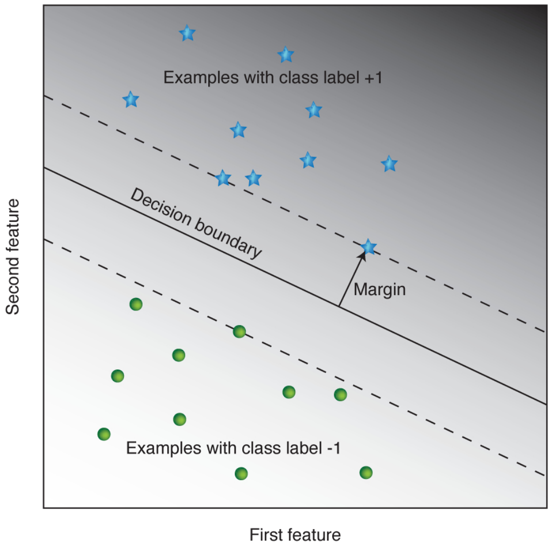


Figure 1.1: Classification with a linear large-margin separator. The decision boundary separates the space (shown for two features) into two half-spaces. Depending on the half in which examples are located, they are classified into different classes represented by stars and circles. The margin is defined as the distance between the decision boundary and the closest example. Among all linear decision boundaries that correctly separate the examples, (hard-margin) SVMs select the one which maximizes this margin.

For more than a decade, *discriminative* machine learning techniques have very actively been investigated. The most well-known representative, the support vector machine (SVM), classifies examples into two classes by means of the *large-margin separation* principle [e.g., 6, 153]. It optimizes a linear decision boundary separating the classes with respect to the distance between the closest examples and the decision boundary. Maximizing this so-called *margin* confers robustness to the classification (Fig. 1.1 for an illustration). The large margin principle has been generalized to other learning scenarios, including regression and label sequence learning. The former problem is addressed by a technique called support vector regression (SVR) [153], whereas hidden Markov support vector machines (HM-SVMs) have been proposed for the latter task [3, 184]. These are related to hidden Markov models (HMMs) [e.g., 44], a *generative* learning method which is ubiquitous in computational biology and which has been applied to various segmentation and label sequence learning problems such as gene finding [e.g., 20]. Recently, HM-SVMs as well as other discriminative machine learning algorithms such as conditional random fields (CRFs) [94] have been engineered to successfully tackle diverse problems arising in computational biology and have been shown to achieve very competitive predictive performance [e.g., 7, 40, 63, 140, 154, 155].

Inherently, supervised machine learning methods are limited to problems for which sufficient amounts of reliably labeled data exist. If this is not the case, training in a purely supervised fashion can be problematic and accuracy assessment impossible or misleading. In such situations, statistical testing, or unsupervised or semi-supervised learning approaches may be preferable. However, large experimental data sets exist for many genomics and transcriptomics problems, such as polymorphism discovery, gene finding, and genome annotation. In these scenarios the application of state-of-the-art supervised machine learning methods is particularly promising.

1.2 Whole-Genome Tiling Microarrays

Whole-genome DNA microarrays are a high-throughput technique that can be seen as a refinement of gene-centric microarrays commonly used in molecular biology and medicine to measure gene expression. In general, DNA microarrays exploit the high specificity with which duplexes between complementary single-stranded DNA (or RNA) molecules form. On a microarray, single-stranded DNA *probes* with known sequence are immobilized or directly synthesized. Currently, high-density microarrays can be manufactured with individual probe features as small as $5\,\mu\text{m}^2$. After hybridizing fluorescently-labeled target nucleic acid sequences prepared from a biological sample of interest (e.g., cellular RNA), bound targets are quantified for each feature by a microscopic imaging device. Probe features on gene-centric microarrays are typically designed to be complementary to known (annotated) genes or expressed sequence tags (ESTs), and therefore only these sequences can be quantified. Tiling arrays, in contrast, interrogate a whole genome (or large genomic region) with probe features *tiling* the (nonrepetitive portions of the) sequence of interest with a regular spacing. The average distance between the centers of tiling probes with adjacent genomic location defines the *resolution* (or step size) of the array.

Due to their design, tiling arrays are a very versatile experimental tool for studying an organism's genome or transcriptome in a manner that is not biased by the current state of its genome annotation. They have been used for experiments as diverse as transcriptome profiling in a global and quantitative manner, chromatin immunoprecipitation ("ChIP on chip") for characterizing transcription factor binding sites, elucidating the histone code or assaying chromatin accessibility, as well as DNA sequence variation detection and methylation mapping [comprehensively reviewed in 117, 200]. In the following I will focus on two applications: global characterization and quantification of transcripts as well as polymorphism detection (also referred to as *array-based resequencing*).

1.2.1 General Properties of Tiling Arrays

Tiling arrays are a massively parallel high-throughput technology that facilitates completing whole-transcriptome measurements within hours (excluding time for sample preparation). Transcriptome tiling arrays are cost-effective in the sense that replicates are usually affordable and the transcriptome can be monitored under a large number of conditions. In contrast, resequencing array experiments are typically not replicated and the number of genotypes assayed rarely exceeds twenty. All tiling arrays produce quantitative data with a large dynamic range, and reproducibility (after appropriate normalization of the raw data, see e.g., Sections 2.2 and 2.3) is usually good despite non-trivial signal noise.

However, all DNA microarrays are limited in the sense that, for their design, target (reference) sequences have to be known beforehand. The analog nature of the intensity signal poses an analysis challenge, particularly so since the relationship between concentration of bound target and signal intensity is complex, e.g., non-linear due to saturation effects. Generally, the intensity signal appears noisy due to (i) experimental variability and failures, e.g., uneven hybridization across the same array and even more dramatically so across different arrays, and (ii) systematic influences on signal intensity other than target concentration, e.g., probe-sequence effects, RNA secondary structure and cross-hybridization [144, 161]. The analysis of microarray data is therefore non-trivial and computational methods almost always apply statistical (learning) techniques. Within the scope of this thesis, preprocessing and normalization methods are proposed that address cross-hybridization issues (see Section 2.1) and reduce systematic noise stemming from divergent sequence properties of oligonucleotide probes that have not been optimized due to constraints on the tiling array design. In particular, we have developed a regression model which estimates deviations between observed and expected hybridization intensities from oligonucleotide sequences. This transcript normalization (TN) method effectively reduces the variance among hybridization signals from different tiling probes which measure the same transcript and are thus ideally expected to produce identical signals (see Section 3.2).

1.2.2 Previous Transcriptome Studies Using Tiling Arrays

Most tiling arrays used for transcriptome studies employ oligonucleotide probes of length 25-70 nt (depending on the manufacturer) with a resolution that varies between 4 and 35 nt. On the *Arabidopsis Tiling 1.0R Array* manufactured by Affymetrix, 25-mer oligonucleotides are spaced at approximately 35 bp intervals, tiling the whole euchromatic genome of *Arabidopsis thaliana*. Upon hybridization with labeled cDNA transcripts converted from cellular RNA, one generally measures strong signals for tiling probes mapped to exons and weak signals for probes mapped to introns and intergenic regions. Probes partially overlapping exons generally produce intermediate signal values (Fig. 1.2).

The first tiling array-based transcriptome studies targeted the *E. coli* genome and human chromosomes [86, 143, 158, 165]. Right from the beginning, emphasis was put on profiling several conditions or cell lines and identifying differentially expressed transcripts. Additionally, a major motivation for tiling array application was to confirm, refine and complete existing genome annotations through the *de novo* detection of expressed transcripts. Surprisingly, these less biased assays of transcriptional activity implied that a much larger fraction of the human genome is transcribed and processed into mature transcripts than previously estimated on the basis of EST data [8, 28, 86, 143, 149]. Evidence of widespread transcription and expression outside of annotated gene and exon boundaries was later also reported for other organisms including *A. thaliana* and *Drosophila melanogaster* [108, 174, 198]. This gap between current annotations and microarray-based

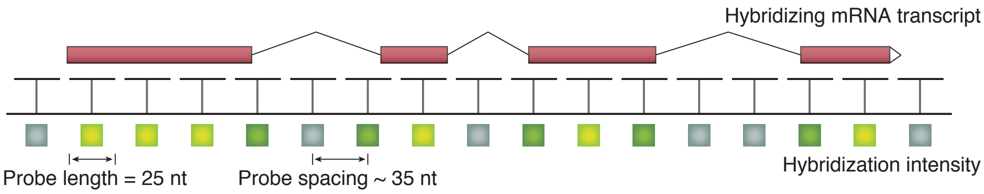


Figure 1.2: Design principles of transcriptome tiling arrays. Oligonucleotide probes are regularly spaced to interrogate the entire reference genome. Transcription can thus be assayed without annotation bias. Indicated probe length and spacing refer to the *Affymetrix Tiling 1.0R Array*, but depending on the design of the array, resolution can be increased until probes overlap. For transcription assays, cellular RNA is converted to cDNA, fragmented and fluorescently labeled before it is hybridized to the array (see also Section 4). As a result, for tiling probes complementary to exons, a hybridization signal proportional to the gene expression level is expected (bright colors indicate high intensity).

transcriptome data prompted speculation on the “dark matter of the genome” [79]. Possible explanations include the incompleteness of the catalog of protein-coding genes, limited knowledge about non-coding and non-polyadenylated transcripts, uncertainties about the prevalence of antisense transcription as well as the extent of alternative splicing and alternative transcription initiation. Consequently, several more recent studies used tiling arrays to characterize certain classes of transcripts, such as RNAs with localization restricted to the nucleus [87], small RNAs with potential regulatory roles [87], non-polyadenylated transcripts [28, 65] as well as noncoding transcripts. In wild-type cells, such non-coding RNAs may be undetectable when they are repressed, degraded or epigenetically silenced. In null mutants of the RNase P enzyme [147], of catalytic exosome subunits [26] or methyltransferases functioning in DNA methylation [208], such non-coding transcripts accumulate and are thus more easily detected.

Searching for explanations of the “dark matter” phenomenon, it was further discussed to what extent tiling array-based transcription data could be the result of experimental artifacts or of false positive computational predictions [79]. Discrepancies found between different tiling array-based analyses for the same regions of human chromosome 22 [86, 143, 149] substantiated these concerns [79]. Although it appeared unlikely that false positive predictions alone were sufficient to explain the phenomenon, it highlighted the need for accurate computational tools. Amazingly, many analyses proceeded in an *ad hoc* manner involving hand-tuned parameters [28, 65, 83, 85, 86, 87, 108, 143, 198], whereas statistically rigorous approaches with a reasonable balance between false positive and false negative rates were (and still are) rare [42, 55, 77, 120, 133].

1.2.3 State of the Art in Transcript Identification from Tiling Array Data

In the following, I will briefly review conceptually related approaches to transcript identification from a single sample while omitting approaches that are only applicable to the identification of regions with significant expression changes between two or more samples [e.g., 77, 133].

The most widely used computational method for transcript identification from tiling arrays is based on a sliding window approach, which, in the first step, detects probes showing significant expression within a local context. In the second step, regions containing several expressed probes interrupted by only a few non-expressed probes are reported as so-called “transfrags” (transcribed fragments), using an *ad hoc* procedure [85]. In contrast to most of the methods reviewed below, this method is non-adaptive and the generated transfrag data strongly depends on user-specified parameters.

For the analysis of *Saccharomyces cerevisiae* tiling arrays, Huber et al. [72] proposed a method that segments the yeast chromosomes such that the sum of the squared differences of signal intensities to their mean within a given segment is minimized. To solve this problem, also known as Structural Change Model Segmentation (SCM), a dynamic programming algorithm was adopted. One advantage of this method is that it very flexibly handles large variance in expression between different genes. However, while this relatively simple approach has been successfully applied to yeast tiling array data, the segmentation problem is considerably more challenging for the transcriptomes of higher eukaryotes, which have less gene-rich genomes and are capable of (alternative) splicing. Note furthermore that it is not an adaptive learning approach and therefore the maximum number of segments has to be specified by the user (to avoid the trivially optimal solution of one segment per probe).

A more sophisticated model, called GenRate, has been proposed by Frey et al. [56]. It explicitly models coregulated units (CoRegs) such as exons that belong to the same gene and hence exhibit the same expression level. However, the generative model for sequences of hybridization measurements, which constitutes the core of their method, is based on several assumptions about the structure of a transcript and the distribution of hybridization measurements (e.g., Gaussian distribution of intensity differences from a designated reference probe, geometrically distributed distance of the reference probe from the transcript start etc.).

Transcript identification has also been approached using HMMs [42, 120, and others]. In contrast to heuristic methods, HMMs offer a principled and flexible inference framework. Although the proposed state models are very different between different HMM-based methods, they all share one limitation (albeit not a conceptual limitation of HMMs): Exons with very different expression levels have to be recognized by the same state. Training such an exon model on genes showing low expression can compromise its ability to distinguish between exon and background signals (see also Section 3.3.2). In Munch et al. [120] this problem is addressed by an additional unsupervised training step subsequent to supervised learning from annotated genes. The unsupervised re-training allows the model to down-weight label information for wrongly annotated genes or genes that are not expressed in the RNA sample hybridized to the tiling array. To the best of our knowledge, none of the previously proposed HMM methods has incorporated an explicit intron model.

To my knowledge, only a single method has been proposed which predicts transcripts from a combination of hybridization signals and features of the genomic DNA [183]. This method, called ARTADE, first identifies probes with significantly higher hybridization intensity than the background using a sliding window approach. In the next step, a Markov model is employed to infer the exon-intron structure of transcripts extending from these highly significant spots of expression. The Markov chain not only models hybridization, but also DNA features such as nucleotide composition.

1.2.4 Key Contributions of This Thesis to Tiling Array-Based Transcriptome Profiling

Within this thesis, a machine learning-based method dubbed **m**argin-based **s**egmentation of **t**iling **a**rray **d**ata (mSTAD) for the *de novo* identification of expressed transcripts is presented (see Section 3.3). It was designed to segment the hybridization signal along a tiling path into intergenic regions, exons and introns. An extension of this method also exploits splice site predictions from the genome sequence as an additional feature in order to more accurately delineate intron boundaries. Our approach is based on HM-SVMs [3, 184], which combine the advantages of HMMs [44] for label sequence learning with those of discriminative SVM classifiers [e.g., 153].

In contrast to the SCM method by Huber et al. [72], we make use of more flexible *scoring functions* replacing the squared error terms in the SCM formulation. Their shapes are estimated from data in order to predict the optimal segmentation of the sequence of intensity measurements. As a supervised learning approach, mSTAD is trained on hybridization intensities together with the correct segmentation determined from known mRNA transcripts.

Our method supersedes previous HMM-based approaches in that it uses a state model that comprises intron states and is thus able to exploit correlated expression between exons of the same transcript (similar to Frey et al. [55, 56]). Additionally, the state model is composed of submodels, each of which is specialized for a certain expression range (in this regard mSTAD can be seen as an adaptive, discrete approximation of the SCM method [72]). These expression-specific submodels can be fitted more precisely, and one thus expects improved accuracy for the recognition of expressed genes. Parameter estimation for such a comprehensive model profits from the discriminative HM-SVM training approach, which optimizes discrimination accuracy between exon and background probes rather than learning a generative model of hybridization intensity. The fact that HMMs and HM-SVMs are closely related inference techniques allowed us to comprehensively compare their predictive performance in the task of transcript identification from tiling array data (see Section 3.3.2). We showed that transcript recognition accuracy is significantly improved compared to the most widely used competing method, an *ad hoc* approach to identify transcribed fragments ("transfrags") [85], and could be even further enhanced when the intensity data were first preprocessed with our transcript normalization method (see Section 3.2.5). The mSTAD framework was readily extended to incorporate splice site predictions as an additional feature to obtain transcript structures with significantly improved accuracy, especially for intron predictions (see Section 3.4.4).

Analyzing data from the Affymetrix GeneChip© Tiling 1.0R Array with mSTAD, we identified thousands of genomic regions that give rise to transcripts expressed in diverse tissues or under stress conditions and that have previously escaped detection by cDNA cloning, EST sequencing or gene prediction approaches (see Section 3.3.3). With RT-PCR experiments, we achieved a validation rate of more than 75% for a small sample of newly predicted transcripts representing diverse lengths and expression level, which further corroborates mSTAD's high accuracy. For the benefit of the research community, mSTAD's source code as well as example data are freely available.

1.2.5 Previous Work on Array-Based Resequencing

Describing the complement of sequence variation within a species is the first step in linking genetic variation to phenotypes [31], and the development of methods for whole-genome polymorphism discovery has been a top priority in the life sciences [162]. Towards this goal, the creation of high-density oligonucleotide microarrays suitable for whole-genome variation detection was a major technological breakthrough [e.g., 25, 68, 131]. Within a decade, these so-called resequencing arrays had scaled from 135,000 features interrogating the human mitochondrial genome [25] to roughly 1.5 billion features complementary to about 58% of the mouse genome [54].

Resequencing arrays employ a 1-bp tiling path to query sequences relative to a known reference sequence. Each base is interrogated with eight features that consist of forward and reverse strand 25-mer *probe 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 perfectly matching (PM) oligonucleotide, thereby predicting the base in the corresponding target DNA sample (Fig. 1.3, Fig. 1.4 A).

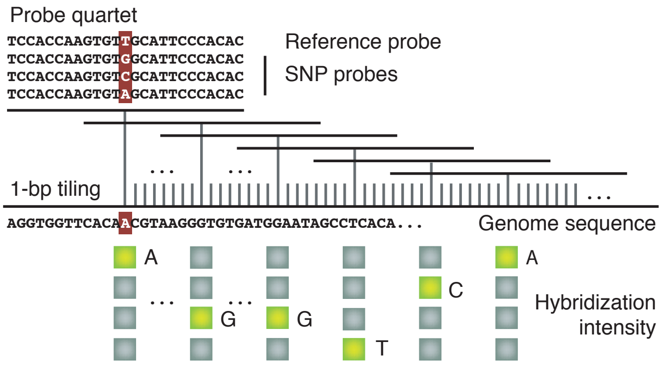

Probe quartet
TCCACCAAGTGTTGCATTCCCACAC
TCCACCAAGTGTGGCATTCCCACAC
TCCACCAAGTGTCGCATTCCCACAC
TCCACCAAGTGTAGCATTCCCACAC
Reference probe
SNP probes
1-bp tiling
...
...
...
AGGTGGTTCACAACGTAAGGGTGTGATGGAATAGCCTCACA...
Genome sequence
A
...
G
...
G
T
C
A
Hybridization
intensity

Figure 1.3: Design principles of resequencing arrays. Each nucleotide of the reference genome sequence is interrogated with probe quartets for the forward and reverse strands (only one quartet is shown). Within each quartet, probes are identical to the reference sequence except at the central position where the reference allele as well as all possible single nucleotide variants are represented. Upon hybridization to labeled genomic DNA, the highest intensity is expected for the perfect match probe indicating the allele in the target DNA sample.

Although conceptually simple, detection of polymorphisms from resequencing array data computationally challenging [29, 34, 131]. For single nucleotide polymorphisms (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 (Fig. 1.4 B). The resulting hybridization pattern provides a "SNP signature", which has been exploited by several algorithms to predict SNPs from resequencing array data [29, 68, 131]. However, where multiple SNPs or insertion/deletion polymorphisms (indels) 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, either in the form of closely linked SNPs or small indels, or potentially of larger deletions (Fig. 1.4 B). This phenomenon has limited the utility of resequencing array data for describing patterns of genome-wide sequence variation. Regions where no SNPs are predicted may be (i) monomorphic to the reference sequence or alternatively, may be (ii) so dissimilar that no underlying polymorphisms are detected.

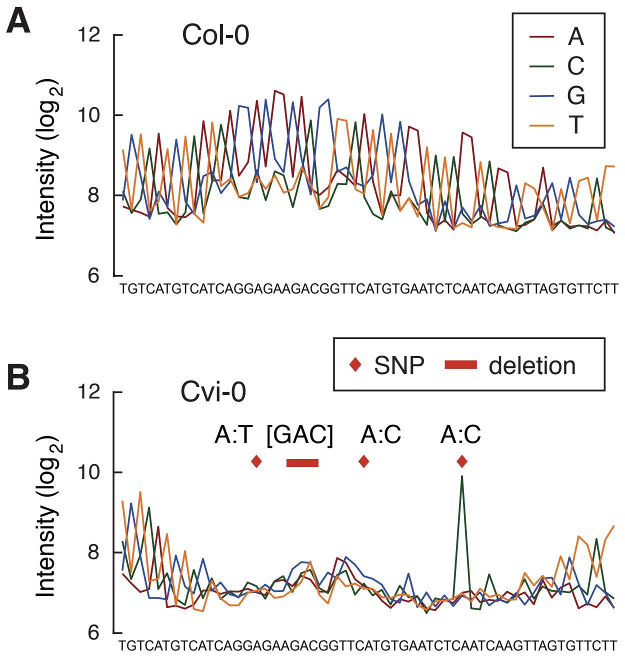


Figure 1.4: Hybridization patterns obtained from resequencing arrays. **(A)** When resequencing non-polymorphic tracts, the maximally hybridizing feature within each probe quartet indicates the base call (see inset). In particular, this is the case for the reference accession Col-0, the genome sequence of which was used for array design. Shown are $\log_2$ intensities averaged across the forward and reverse strand features. **(B)** Polymorphic sequences exhibit a remarkably different hybridization pattern (corresponding data from accession Cvi-0 is shown). 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 3 polymorphisms including the deletion (left center).

Furthermore, advanced computational approaches to the detection of sequence variants other than SNPs, e.g., indels or structural changes (inversions, translocations etc.) from resequencing arrays have not been reported. In one study, Hinds et al. [69] 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 human. More recently, [29, 202] applied a simple heuristic algorithm to predict tracts of highly divergent or missing sequences from similar data for *A. thaliana*. Although this heuristic algorithm generated several hundred predictions per accession, it only identified extended polymorphic tracts (about 300 bp to many kb) consisting largely of deletions. Currently, no methods have been reported to predict short indels (tens of bp) or clustered SNPs from resequencing array data. This limited investment in methods reflects, in part, the complex nature of the primary data [29, 202]. 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 [29, 54, 68].

1.2.6 Contributions of This Thesis to Array-Based Resequencing

Due to the relatively high noise levels in tiling array data, adaptive learning approaches are ideally suited for their analysis. A machine learning method able to predict regions of high polymorphism density from resequencing array data is described in this thesis (Section 3.1). It was developed for the task of labeling each tiled position in the genome as either (i) conserved or (ii) at or immediately adjacent to a polymorphism. This method, which we call margin-based prediction of polymorphic regions (mPPR), employs HM-SVMs [3, 184] modeling the array measurement sequences to learn to identify *polymorphic regions* (PRs). We applied mPPR to an *A. thaliana* resequencing array data set for 20 accessions that contains data generated for $>99.99\%$ of bases in the 119 Mb reference genome [73] for each accession [29] (see Section 3.1). This data was previously used to identify approximately 648,000 SNPs at a precision of about 98% [29]. With mPPR, on average about 288,000 polymorphic regions were predicted per accession at a precision of about 97%. Non-redundantly, 27% of the genome was included within the boundaries of PRs. A large proportion (about 66%) of a set of known SNPs were contained in PR predictions, of which 42% were absent from the previous SNP data set. Because of its applicability to similar data sets in other species, our method has already been utilized to characterize tracts of elevated polymorphism in rice cultivars [10] (see Section 3.1).

1.2.7 Alternative Technologies to Tiling Microarrays

Technologies competing with tiling arrays for addressing similar genomics and transcriptomics questions are mostly based on DNA sequencing.

Alternative Approaches to Genome Resequencing

Traditionally, resequencing and sequence variation detection was carried out using the classical dideoxy sequencing method. Despite substantial costs associated with this technology , large genome-wide polymorphism data sets had already been collected before array-based resequencing became feasible for medium to large genomes [31, 128]. However, an array-based resequencing project for human greatly increased the number of SNPs with associated allele frequencies deposited in the dbSNP database [68]. Before Clark et al. [29] published an array-based polymorphism data set for*A. thaliana*, no whole-genome inventory of SNPs had existed for this model plant.

About two years ago, “next-generation” sequencing (NGS) technologies began to revolutionize genome research by dramatically increasing the throughput and lowering the costs for large-scale sequencing efforts. Pyrosequencing (Roche / 454) and sequencing by synthesis (Illumina / Solexa) as well as bead-based sequencing (ABI / SOLiD) have been applied in numerous resequencing projects including one that targets *A. thaliana* accessions (The 1001 Genomes Project[1]) [130, 189]. NGS technology is particularly applicable to resequencing since its major drawback, short read lengths, can be overcome more easily than for *de novo* sequencing. In the resequencing scenario, it is typically sufficient to map

[1] `http://www.1001genomes.org`

the reads to the reference genome sequence in order to very accurately detect the vast majority of SNPs and small insertions (up to 3 bp) as well as deletions [e.g., 130]. However, for resolving other structural variants, such as large insertions, assembly algorithms addressing the core problem of *de novo* shotgun sequencing are inevitably needed. Although assembly is extremely challenging for small reads ($\sim 35\,\mathrm{nt}$), it does facilitate detecting at least a fraction of structural variants [e.g., 130]. In the near future, genome resequencing will greatly benefit from extensions of sequencing by synthesis in the direction of paired-end reads. This additional information will likely suffice to resolve structural variants in great detail and depth [91]. Moreover, latest advances in single-molecule sequencing (one of the emerging technologies that are sometimes called "next-next generation sequencing") indicate that it may be possible to routinely obtain sequence reads that are several kb in length [47]. This would open up entirely new perspectives for resolving structural variants and repetitive sequences — in the context of resequencing as well as for *de novo* genome sequencing.

Alternative Approaches to Transcriptome Profiling

Transcriptome studies have for a long time been performed using EST-sequencing techniques [1]. However, obtaining deep coverage is very expensive and the completeness of EST-based genome annotations has therefore been challenged by many transcriptome studies using tiling arrays [8, 28, 86, 102, 103, 108, 143, 149, and others]. For the same reason, EST-sequencing does not allow a reasonable quantification of transcripts. Thus, to be able to quantitatively assay the transcriptome with sequencing-based methods, tag-based / signature sequencing methods have been developed [reviewed in 64]. Here, only small fragments ($\leq 20\,\mathrm{bp}$ tags) from one or both ends of transcripts are sequenced, which increases throughput into the range needed for quantification. However, focusing the sequencing onto small tags inevitably results in sparse and non-uniform coverage of transcripts. Consequently, these tag-based approaches are better suited to confirm predicted transcripts than to identify transcript structures *de novo*. Lately, the first applications of NGS technology to transcriptome sequencing (or "RNA-seq" for short) have been published [110, 118] [reviewed in 161, 188]. RNA-seq data has been shown to cover a large range of transcript quantities, and the digital sequence data makes it possible to accurately quantify and resolve (even alternative) transcript structures. Yet, data analysis is challenging, owing mainly to the small read length. Also, read coverage is highly non-uniform even within the same transcript and reproducibility has not been tested so far [reviewed in 161, 188]. Although RNA-seq data and tiling array data differ with respect to many characteristics, our transcript normalization method may also be applicable to RNA-seq data. Reducing sequence biases, which this kind of data exhibits to a similar extent as that of tiling arrays, may improve transcript deconvolution and quantification from RNA-seq data (see Section 3.5.2).

1.3 Arabidopsis thaliana as a Model for Studying Natural Sequence Variation

The flowering plant *Arabidopsis thaliana* is a small weed from the mustard family (Fig. 1.5 A), which also contains some members of agricultural interest, e.g., cabbage, cauliflower, rapeseed, radish, and turnip. Although *A. thaliana* itself is not an economically significant plant, it was one of the first multicellular model organisms for which a reference genome sequence became available, when the 120 Mb genome sequence of the Col-0 accession was finished in 2000. While this was a milestone for plant research, it has become increasingly clear that "a single genome is not enough" [189]. Therefore, past and future sequencing efforts have aimed to characterize sequence variants in a large number of *A. thaliana* accessions [29, 128, 130]. Sequencing the genomes of many individuals to complement a reference sequence is being carried out in human [101]. Tremendous efforts are under way [84] towards creating a new map of biomedically relevant human DNA variations at previously unmatched resolution with the aim of providing the foundation for personalized medicine. Projects for *A. thaliana* and *Drosophila melanogaster* are proceeding in parallel [189],[2] and are advancing fast, owing to the smaller genome sizes. These model organisms are resequenced with the primary goal of establishing comprehensive polymorphism resources, which are extremely valuable for studying comparative genomics, system genetics and molecular evolution. Fine-scale conservation information will, for instance, be leveraged to pinpoint functional DNA elements. Moreover, high-resolution sequence variation data is the basis for genome-wide association studies aiming to genetically map complex traits [127].

A. thaliana has a wide distribution over large areas of the northern hemisphere, and accessions isolated from natural populations exhibit enormous phenotypic variation in morphology and physiology (Fig. 1.5 B and [189]). Because *A. thaliana* proliferates predominantly by self-fertilisation, isolates from natural populations are largely homozygous making them particularly amenable for polymorphism discovery and genotyping. In contrast to humans, individual plants that are genetically identical to the ones genotyped are available to the research community and can be phenotyped in a large number of environments. *A. thaliana* is thus an ideal model for investigating how genetic variation causally relates to complex (quantitative) phenotypic variation and adaptation [see, e.g., 189].

1.3.1 Contributions of This Thesis to Revealing Patterns of Polymorphisms in Arabidopsis

The whole-genome resource of polymorphic regions predicted by mPPR with a false discovery rate of <3% contained between ~240,000 and ~361,000 PRs per accession [203]. These provided a fine-scale view of polymorphic sequences in *A. thaliana* allowing us to confirm findings that were initially based on SNPs identified from the same array data: Polymorphism levels are very non-uniform across gene families, and NB-LRR genes, which are involved in disease resistance, show extremely high levels. PR data is consistent with

[2] `service004.hpc.ncsu.edu/mackay/Good_Mackay_site/DBRP.html`

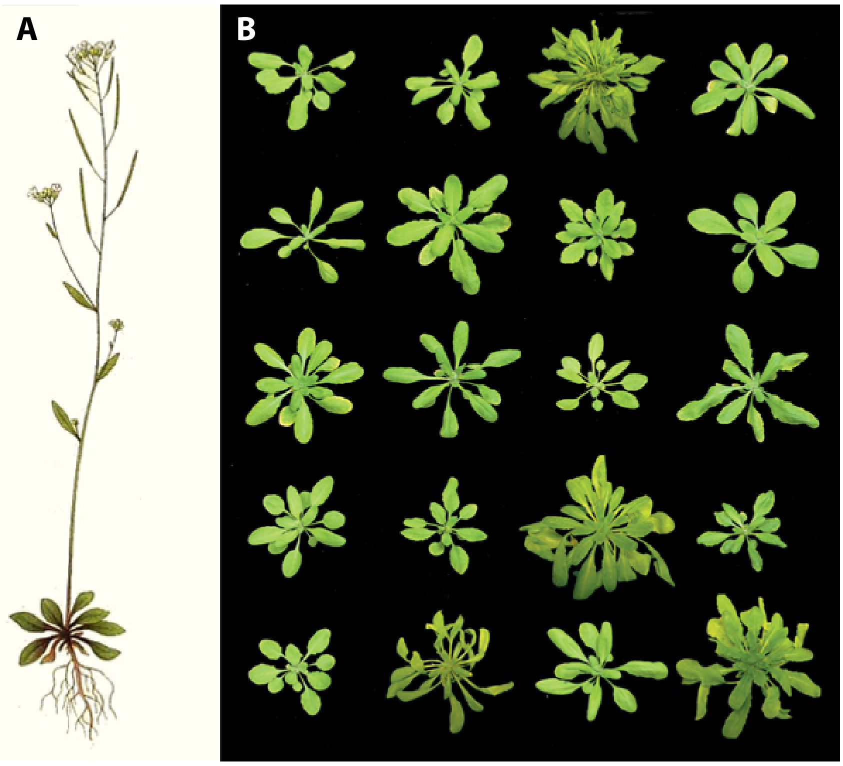


Figure 1.5: The model plant *Arabidopsis thaliana.* **(A)** Drawing of a flowering *A. thaliana* plant from *Carl A. M. Lindman's Flora* (digital image processed by Project Runeberg, Dr. Gerhard Keuck). **(B)** Morphological variation among different *A. thaliana* accessions grown under the same environmental conditions (courtesy of Detlef Weigel's laboratory).

previous publications [4, 160] reporting that many members of this family are deleted or completely divergent in some of the accessions, and we experimentally confirmed one such case. Also, several members of the F-box genes [see also 181] and receptor-like kinases harbor extreme polymorphism levels. However, as a whole, these gene families exhibit less extreme distributions (see Section 3.1.5). Although striking polymorphism patterns have been described previously for many genes [4, 60, 61, 160, 166], our inventory of polymorphic regions provides, for the first time, a comprehensive, fine-scale view of the whole *A. thaliana* genome.

Additionally, the PR data revealed patterns of polymorphism not apparent in SNP data, especially for noncoding regions, e.g., around gene and exon boundaries and in miRNA precursor genes. Generally, patterns of intra-specific sequence variation resembled those observed for inter-specific variation: Exons have lower polymorphism levels than introns; in miRNA precursor genes, levels were lowest in the regions that are processed into the mature miRNA and miRNA*; and in core promoters, reduced levels were found in predicted cis-elements and in the TATA-box (see Section 3.1.4).

1.4 Arabidopsis thaliana as a Model for Transcriptomics

Being extensively studied as a model for genetics and developmental biology, *A. thaliana* has been instrumental in gaining insights into plant development and physiology. This includes, for instance, flower formation, photosynthesis, plant biomass production, root development and processes governing the uptake of water and nutrients. Discoveries made in *A. thaliana* have even impacted human health [reviewed in 80]. In comparison to economically important crop plants, *A. thaliana* is more easily grown and genetically manipulated, and therefore experimental and computational resources for *A. thaliana* are unmatched among plant species. Although existing gene annotations are of high quality [178] and supported by extensive experimental efforts [e.g., 157, 198], they are still being continuously improved. Genome-wide expression maps, which complement these static annotations, have become an indispensable tool for the research community [e.g., 9, 89, 124, 152, 171] that has helped to elucidate transcriptional networks and attendant promoter motifs, to uncover gene functions and to reveal molecular explanations for mutant phenotypes [reviewed in 21].

Plants are also a prime model for studying stress response, with important implications for agriculture. As limiting factors for agricultural productivity, abiotic stresses — salt and drought in particular — have been a focus of research. Understanding stress response in the model plant *A. thaliana* may also help to increase the tolerance of crop plants to adverse environments and climate change. Eventually, this might increase (or at least stabilize) yield and the proportion of arable land [reviewed in 38]. Being sessile, plants cannot move away from extreme conditions such as heat, cold, high salinity or drought. These stress situations trigger signals that alter plant physiology and growth to ensure survival in hostile environments. Ultimately, signaling cascades downstream of stress sensors result in altered expression of stress-responsive genes. Some of these encode proteins responsible for the biosynthesis of hormones, such as abscisic acid (ABA), which can act as signaling molecules that amplify and spread the initial stress signal. Interestingly, different stresses as well as ABA treatment can change the expression of a common set of genes, indicating that stress responses are mediated in part by overlapping signaling pathways [e.g., 75, 199]. However, these common signaling pathways might be activated in a different temporal and spatial manner by individual stresses [e.g., 39, 89, 197]. In addition, there are signaling events that are specific to a particular stress [199, and references therein]. Together, these differential responses enable the plant to react adequately and specifically to different stresses.

Several reports have characterized transcriptome changes in plant organs, during development, and under abiotic stresses [9, 27, 50, 89, 89, 92, 124, 152, 156, 171, among others]. However, the vast majority of expression analyses have been performed with full-length cDNA arrays or oligonucleotide arrays targeting known transcripts. The main disadvantage of these techniques is that they rely on prior information about potentially transcribed regions based on cDNA cloning, ESTs or computational gene predictions. The most widely used microarray platform for expression analyses in *A. thaliana* is the Affymetrix ATH1 array, first launched in 2002 [141]. Its design was based on experimentally confirmed transcripts and gene predictions, but now lacks almost a third of the 32,041 genes found in the TAIR7 annotation [178]. All users of ATH1 arrays are thus confronted with this problem: As the number of newly discovered genes rises, expression analysis becomes more and more restricted. These limitations may be even more severe for the analysis of stress-induced expression changes, because transcripts appearing only under extreme environmental conditions may well have escaped previous annotation efforts. Especially in the light of the growing appreciation of the roles of non-coding RNAs, a more unbiased detection of stress-induced changes of the *A. thaliana* transcriptome is of great importance [e.g., 51, 107, 177]. Currently, whole-genome tiling arrays are among the most cost-effective approaches to more unbiased detection of transcriptional activity. Creating a tiling array-based community resource for *A. thaliana* transcriptomics was thus very timely.

1.4.1 Contributions of This Thesis to Arabidopsis Transcriptomics

Having analyzed *A. thaliana* tiling array data for eleven tissues and developmental stages [97] and, additionally, data for five abiotic stresses collected at two time points after treatment [205], we created an initial whole-genome expression atlas, dubbed **A***rabidopsis* **t***haliana* **t**iling **a**rray e**x**pression atlas (At-TAX)[3]. It is intended to complement or even replace gene expression atlases for *A. thaliana* that are based on conventional gene-centric ATH1 microarrays, such as AtGenExpress [152]. In addition to expression estimates for a much more complete gene catalog than could be surveyed with ATH1 arrays, At-TAX visualizes transcriptionally active regions (TARs) identified with mSTAD, including between 925 to 1,947 new TARs per sample that were predicted with high confidence, but had been overlooked by previous annotation efforts (see Section 3.3.3). Several hundred (84 to 375) of these show significant expression changes under stress, and for a selected set, we could experimentally confirm differential expression (see Section 3.3.4). Subsequent analyses suggest that several of these result from new genes, whereas the majority may either be new constitutive exons of annotated genes or belong to stress-specific alternative transcript isoforms arising from known loci (see Section 3.3.4). A comparison of TARs generated from tiling arrays assaying polyadenylated RNA or total cellular RNA suggested that the majority of *A. thaliana* transcripts that are expressed at detectable levels are polyadenylated (see Section 3.3.3).

[3] `http://www.weigelworld.org/resources/microarray/at-tax`

2 Methods

This chapter first describes methods for the identification of repeats that are most likely to produce artifacts in microarray analyses. Subsequently, properties of resequencing and transcriptome tiling arrays are presented together with appropriate normalization techniques. The central sections are dedicated to the HM-SVM learning algorithm, an investigation of its basic properties in comparison to HMMs and its adaptation to the bioinformatics problems of polymorphic region prediction and transcript identification from tiling arrays.

2.1 Detection of Repetitive Oligonucleotide Probes

As for many other high-throughput technologies, genomic sequence repeats pose a challenge for the interpretation of microarray data. Hybridization signals from probes on the microarray that are complementary to genomic repeats may result from cross-hybridization and may thus reflect a complex mixture of hybridizing targets originating from several places in the genome. Attributing such a signal to a unique genomic location is therefore impossible and deconvoluting cross-hybridization signals is generally very difficult.

To avoid cross-hybridization artifacts in downstream analyses, we therefore determined for each array probe whether its sequence matches with high complementary to additional genomic locations. We subsequently used this information in the algorithms described below or for *ad hoc* curation of predictions.

2.1.1 Annotating Repetitive Probes on Resequencing Arrays

Cross-hybridization of repetitive sequences confounds polymorphism detection from oligonucleotide arrays, and can either (i) mask legitimate polymorphisms or (ii) introduce anomalous intensity readings for nonpolymorphic regions that lead to spurious polymorphic predictions.

Exact, Short, and Inexact 25mer Matches

We distinguish three classes of matches between repetitive 25mer probes, each of which is allowed a mismatch at the central (13th) position that varies as part of the array design (Fig. 2.1). First, exact 25mer matches correspond to probes that are completely complementary to at least two genomic locations (on either genomic strand) for positions 1-12 and 14-25. Second, because mismatches at the ends of probes have comparatively little effect on hybridization strength [99], we identified short 25mer matches according to the same rules except that mismatches were allowed on any or all of the 2 bp on either end of 25mer probes. Finally, inexact 25mer matches correspond to probes that have multiple complementary counterparts in the genome with one mismatch at positions 1-12 or 14-25. For inexact matches, the potential for stable duplex formation (and for cross-hybridization on arrays) is more difficult to predict, and is expected to vary depending on sequence properties and mismatch location within the probe [99].

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

The sorting algorithm was then modified to handle mismatches. We used a recursive, position-wise partitioning method that begins by partitioning the tiling list according to the nucleotide at position 1 of each 25mer. This partition is then recursively subdivided according to subsequent positions. Mismatches at the central 25mer position are tolerated by skipping the 13th partitioning step. Partitions created when sorting on position 12 are therefore subdivided according to the nucleotides at position 14. The generalization of the sorting method to short 25mer matches is straightforward: in addition to position 13, positions 1, 2 and 24, 25 are skipped.

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

As 25mers had been tagged with genome locations, mapping final partition blocks back to the genome was straightforward. Counts of positions with exact, short, and inexact 25mer matches are given in Table 4.1 and Fig. 4.1. More details on the annotation of repetitive resequencing array probes can be found in Zeller [202].

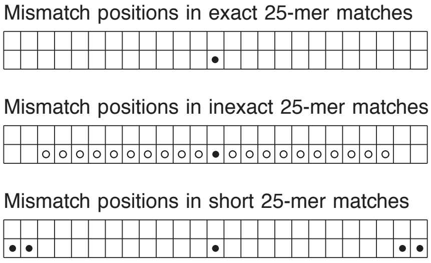

Mismatch positions in exact 25-mer matches
Mismatch positions in inexact 25-mer matches
Mismatch positions in short 25-mer matches

Figure 2.1: Match type definition for 25mers. Squares denote positions in probes from 1 to 25, and filled circles indicate positions at which mismatches are tolerated. For inexact matches, a single mismatch at one of the positions indicated by open circles is tolerated.

2.1.2 Annotating Repetitive Probes on Tiling Arrays

For the annotation of perfect match (PM) probes on the tiling array, we extracted the subset of repetitve resequencing probes with the following restrictions: (i) Only if the reference probe of a quartet of resequencing probes matched without a center mismatch, was the corresponding PM tiling probe annotated as repetitive. (ii) Repeat annotation of resequencing probes was restricted to the subset of positions represented on tiling arrays. (iii) Matches to organellar genomes were excluded, as contamination with organellar genomes was found to be a problem of the resequencing data, but to a much lesser extent for transcriptome tiling array data.

After filtering according to the above criteria, we retained 163,757, 236,832 and 163,757 PM tiling probes with exact, inexact and short matches, respectively. The resulting probe annotation, in which approximately 12.8% of all PM tiling probes (389,264) were associated with any of the above repeat types, was used as a filter for subsequent analyses (see, e.g., Section 3.3.3).

2.2 Resequencing Array Data

Resequencing array data were generated for 20 accessions (derived from wild strains) of *Arabidopsis thaliana* including the reference accession Col-0 by Perlegen Sciences[1] (as described in the supplement of [29]). Due to the costs associated with resequencing arrays interrogating $>99.99\%$ of the euchromatic sequence of the *Arabidopsis thaliana* genome using $>950{,}000{,}000$ different oligonucleotide probes distributed across five large wafers, replicate experiments were not performed. (For more details on the resequencing data see [29, 202].)

2.2.1 Correcting for Between-Array Variability

The raw resequencing array data were characterized by a high degree of variability in hybridization intensity (Fig. 2.2; [29, 202]). To facilitate comparisons between data from different accessions and from different wafers, raw hybridization intensities were quantile-normalized on a per-wafer basis [11]. Quantile normalization involves computing the mean over the empirical intensity distributions from individual wafers. This mean distribution is then re-assigned to each of the wafer distributions, thus effectively removing differences between intensity distribution of different wafers [11]. The extent to which this procedure reduced differences in the intensity distributions of probes interrogating a subset of positions across all wafers corresponding to regions where polymorphisms have been characterized before (see Section 3.1.1; [128]) is visualized in Fig. 2.2. When applying quantile normalization, we assume that differences in hybridization intensities are largely due to experimental variability and not the outcome of a biological process of interest. Although the "true" intensity distributions are probably not identical across accessions, since polymorphism levels are not identical either, we expected the disadvantage of smoothing out

[1] `http://www.perlegen.com/`

some real signal by quantile normalization to be outweighed by the benefit of intensity data comparable across wafers and accessions.

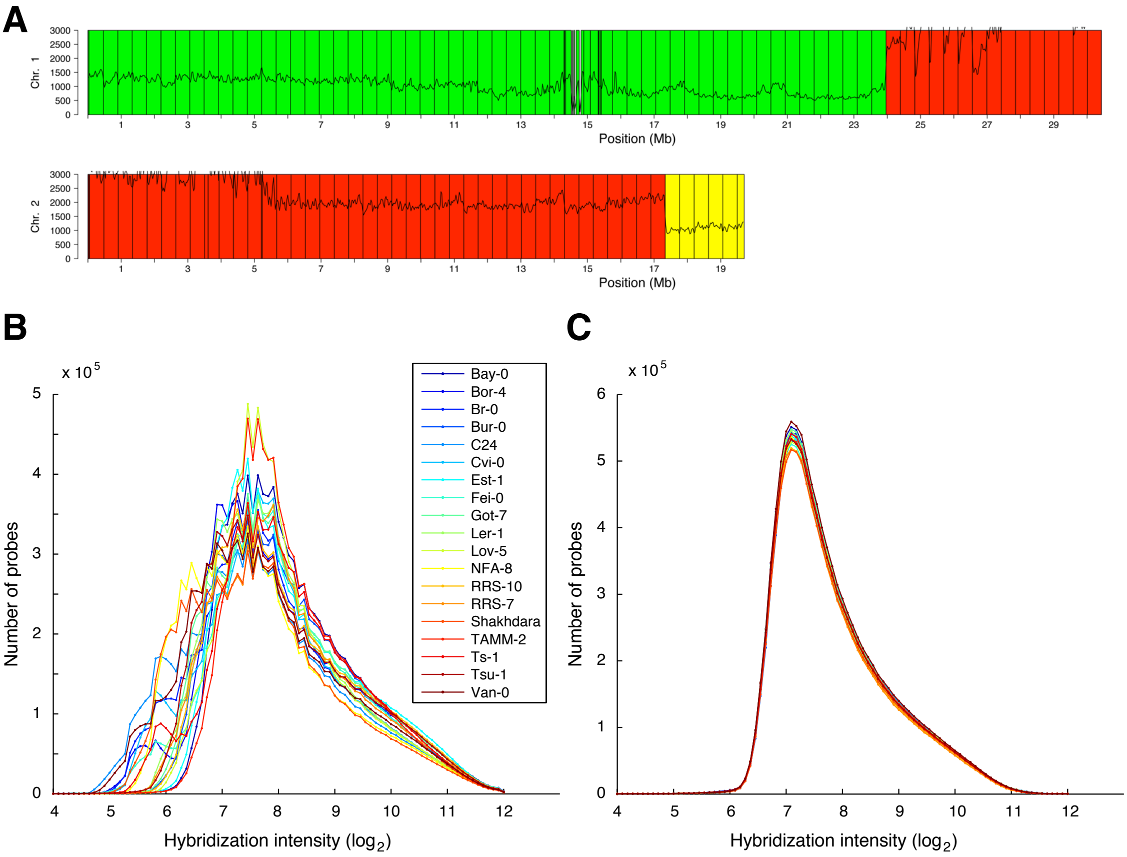


Figure 2.2: Illustration of variability in raw resequencing intensity data and the result of quantile normalization [11] (Reproduced from Zeller [202]). **(A)** Raw intensity data for the accession Bur-0 plotted along chromosomes 1 and 2 in a sliding window of length 15 kbp. Wafers are shaded in different colors and array boundaries are indicated by vertical lines. Note the pronounced discontinuities at wafer boundaries. **(B)** Histograms of raw intensities by accession (see inset) generated from all resequencing probes inclusive to 2010 regions (see Section 3.1.1; [128]) **(C)** Intensity histograms generated from the same probes as in **(B)**, but after quantile normalization [11].

2.3 Transcriptome Tiling Array Data

All tiling array data analyzed in the scope of this thesis were generated with *Arabidopsis* Tiling 1.0R arrays manufactured by Affymetrix[2].

For hybridization, total RNA of whole plants or plant organs was extracted (Table 4.4 for a description of sample generation). Typically, RNA was amplified using oligo-dT-T7 primers which target the polyA tail of polyadenylated messenger RNA molecules (also, some data based on random primers was analyzed and compared to oligo-dT amplified targets in Section 3.3.3). Resulting RNA was converted into double-stranded cDNA,

[2] `http://www.affymetrix.com`

fragmented, labeled and hybridized to Affymetrix tiling arrays (see Section 4 for details of the hybridization protocol).

Prior to data analysis aiming at quantification of gene expression, discovery of new genes or the detection of transcript isoforms, raw array data were subjected to a preprocessing pipeline detailed below. It was established with help from Timo Sachsenberg, Stefan R. Henz and Gunnar Rätsch.

2.3.1 Normalizing for Uneven Background

Background correction of the raw array data aims at reducing spatial variations in the scanned array image that are the result of non-uniform hybridization, washing across the glass slide of the array or spatial scanner and imaging biases. Since spot locations are randomized, i.e., neighboring probes in a tiling path are very unlikely to be spotted next to each other on the microarray, large spatial differences are very unlikely to be the result of a biological phenomenon. Therefore, background correction is routinely applied in standard expression analysis software [74, 194]. The background correction we employed was proposed by Borevitz et al. [14]. It essentially computes background intensity by a sliding window averaging procedure (we used a 51 by 51 feature window) similar to a mean filter commonly used in image analysis. Assuming that background noise is additive [43, 194, among others], this background image is subsequently subtracted from the raw intensities [14].

2.3.2 Correcting for Between-Array Variability

To facilitate inter-array comparisons, all arrays included in subsequent comparative analyses (typically all arrays of the same series) were jointly subjected to quantile normalization [11]. In the context of tiling arrays, his procedure assures that measurements from different arrays but with the same rank (within the respective array) are assigned the same intensity value by quantile normalization. All intensity measurements were then $\log_2$ transformed for the subsequent normalization and analysis steps.

2.3.3 Normalizing for Probe Sequence Biases

Both sensitivity and specificity of binding at DNA microarray probes vary considerably depending on the sequence of the probe itself as well as hybridization conditions [66, 113, 122, 145, 207, among others], and similar issues play a role for primer design [e.g., 109, 146]. In contrast to PCR, hybridization conditions cannot be optimized for individual probes but only for the whole microarray, and therefore probe selection algorithms have been developed, particularly for gene expression arrays, which attempt to select probes with favorable sequence properties given global hybridization conditions [e.g., 113]. However, when probe lengths are restricted to $\approx 25\,\text{nt}$ [132] and probe positions are highly constrained due to the tiling design, binding properties will inevitable exhibit high variability between different tiling probes. Hence, instead of selecting optimal probes, one can only correct the hybridization data *a posteriori* to alleviate probe sequence-dependent biases [66, 145, 207]. Ideally, such a probe-sequence normalization would not require additional hybridization experiments with a genomic DNA control [35, 72], but estimate hybridization bias from the probe sequence itself.

Sequence Quantile Normalization (SQN)

Sequence quantile normalization (SQN) has been proposed as an extension of the above-described quantile normalization to remove probe sequence effects [145]. For each 25mer probe having nucleotide $j \in A, C, G, T$ at position $k = 1, \ldots, 25$, the rank $r_{i,j,k}$ of its intensity y_i among all other probes with the same nucleotide at position k is calculated and normalized by the number of such probes $C_{j,k}$. These position-wise contributions are then averaged: $\hat{S}_i = \frac{1}{25} \sum\limits_{k=1}^{25} \frac{r_{i,j,k}}{C_{j,k}}$. Since the sequence bias is not uniform across positions and summands are not independent, the multivariate regression problem is solved iteratively. In each step, the above average is computed and afterwards intensities y_i are replaced by $\hat{S}_i$ which is repeated until convergence [145].

As a side effect, intensities are substituted by relative ranks which are uniformly distributed between zero and one. In order to obtain normalized intensity values comparable to the original measurements from the array, we modified the averaging as follows. Intensity distributions were approximated by piece-wise linear functions $g_k(r_{i,j,k}) \approx y_i$. In our case g was parametrized by 200 supporting points with uniformly spaced x-values s_x between zero and one. The corresponding y-values s_y were estimated by linear interpolation between y_m and y_n with ranks $r_{m,j,k} = \max\limits_{m'}\{r_{m',j,k} \mid r_{m',j,k}/C_{j,k} \leq s_x\}$ and $r_{n,j,k} = \min\limits_{n'}\{r_{n',j,k} \mid r_{n',j,k}/C_{j,k} \geq s_x\}$, respectively. Instead of averaging relative ranks, we then calculated the mean $\hat{g} = \frac{1}{25} \sum_{k=1}^{25} g_k$ of the supporting points s_y. From this averaged $\hat{g}$ we reconstructed the normalized intensities by linear interpolation between supporting points of $\hat{g}$.

Although SQN effectively reduces probe sequence biases (Fig. 3.13 B and Fig. 3.14), we decided to not include it in our preprocessing pipeline due to its property of increasing within-gene variability. Instead, SQN was replaced by the transcript normalization method detailed below.

Transcript Normalization (TN) Techniques

Ideally, one would expect constant hybridization intensity for all probes measuring the same transcript. Similarly, the background signal of probes in untranscribed or intronic regions of the genome would ideally be constant. However in practice, this is generally not the case [see, e.g., 144, for discussion]. Assuming that the main reason for this discrepancy are probe sequence biases, we developed a new method to effectively reduce the observed within-gene variability. This was joint work with Stefan R. Henz, Sascha Laubinger, Detlef Weigel and Gunnar Rätsch (see p. 137 for author contributions) [204].

In a first step we estimated constant transcript and background intensities $\overline{y}_i$ based on the TAIR7 annotation [178], in the following simply referred to as *transcript intensities*:

For a probe i annotated as exonic, we set $\overline{y}_i$ to the median of the intensities y_i of probes in exons of the same gene. Similarly for intron probes, we computed $\overline{y}_i$ as the median over intronic probes of the same gene and for intergenic regions $\overline{y}_i$ as the median of all probes mapped to regions annotated as intergenic (Fig. 2.3, see Section 2.3.4 for details of tiling probe annotation).

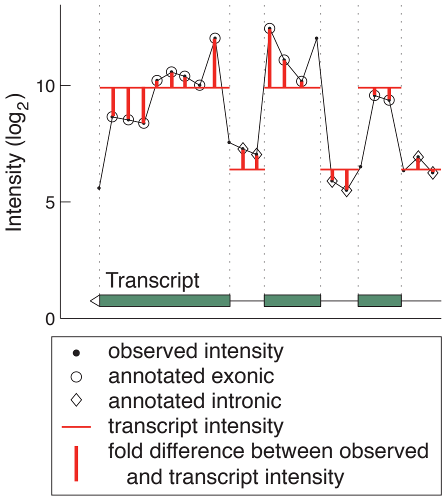


Figure 2.3: Illustration of transcript intensity. Although ideally expected to be constant, hybridization intensity exhibits high variability across all probes complementary to the same transcript. Median estimates of constant transcript intensity as well es the deviation of observed intensities from this constant are shown in red (see inset). Unambiguous probe annotations are indicated by circular or diamond shapes (see inset).

first, we assumed that the concentration of mRNA hybridized to all exon probes of a gene is constant thereby ignoring alternative transcripts. Our second assumption is that the differences between the raw intensities and the transcript intensities $\hat{y}_i := y_i - \overline{y}_i$ are largely due to probe sequence-specific effects — ignoring cross-hybridization, experimental artifacts and thermodynamic noise (Fig. 2.3). Furthermore, it is conceivable that probe effects also depend on the mRNA concentration, and hence the differences $\hat{y}_i$ may also depend on the transcript intensity $\overline{y}_i$ (of the exons) of the gene. Since it is not obvious how this dependency should be modeled, we would like to non-parametrically model the difference by a function that depends on both, sequence features $\boldsymbol{x}_i$ of the probe as well as its transcript intensity, i.e., to estimate a function $f(\boldsymbol{x}_i, \overline{y}_i) \approx y_i - \overline{y}_i$. However, in order to use this correction, one would have to know in advance whether a certain probe is exonic, intronic or intergenic, which is not generally the case. We therefore decided to estimate the function depending not on the transcript intensity, but instead on the raw intensities as a proxy for the former, i.e., $f(\boldsymbol{x}_i, y_i) \approx y_i - \overline{y}_i$.

The large amounts of available data for estimating $f(\boldsymbol{x}, y)$, allowed us to discretize the parameter y into Q quantiles, and subsequently we estimated Q independent functions $f_q(\boldsymbol{x})$. Then $f(\boldsymbol{x}, y)$ had the following form.

$$f(\boldsymbol{x}, y) = \begin{cases} f_1(\boldsymbol{x}) & \text{for } y \in (-\infty, y_1) \\ \cdots & \cdots \\ f_i(\boldsymbol{x}) & \text{for } y \in [y_{i-1}, y_i) \\ \cdots & \cdots \\ f_Q(\boldsymbol{x}) & \text{for } y \in [y_{Q-1}, \infty) \end{cases}$$

As input $\boldsymbol{x}_i$ to the regression function f_q the sequence $\boldsymbol{s}_i$ of probe i was provided together with additional features derived from the sequence: a) sequence entropy $-\sum_{i=1}^{4} f_i \times$

$\log(f_i)$, with f_i being the frequency of the nucleotide $i \in \{A, C, G, T\}$ in the probe sequence; b) GC content; c) two hairpin scores: First, the maximum number of base pairs over all possible hairpin structures that a probe can form. The second one captured the maximum number of consecutive base pairs over all possible hairpin structures (similarly used for intensity modeling in Zhan and Kulp [206]).

Based on these sequence features, we considered two methods for learning the functions f_q based on Q sets of n training examples $(\boldsymbol{x}_i^q, \hat{y}_i^q)$, where $\hat{y}_i = y_i - \overline{y}_i$, $i = 1, \ldots, N$ and $q = 1, \ldots, Q$: (i) Support Vector Regression [153] and (ii) Ridge Regression [70], both of which will be introduced briefly.

Support Vector Regression (SVR) We applied Support Vector Machines [e.g., 153] for regression, employing a kernel function $k(\boldsymbol{x}, \boldsymbol{x}')$ which computes the "similarity" of two examples $\boldsymbol{x}$ and $\boldsymbol{x}'$. Here we used a sum of the so-called "Weighted Degree" (WD) kernel [136, 139, 167, 170] and a linear kernel. The WD kernel has been successfully used to model sequence properties taking the occurrence and position of substrings up to a certain length d into account [139, 140, 155, 167, 169, 170]. We considered substring lengths of up to order $d = 3$ and allowed a shift of up to 1 bp between the positions of the substrings [139], which could be efficiently dealt with using string indexing data structures [168]. The linear kernel computes the scalar product of the sequence derived features described above. These kernels are implemented in the *Shogun toolbox*[3] [168].

Ridge regression (RR) For every training example, we explicitly generated a feature vector from the sequence $\boldsymbol{s}$ having an entry for every possible mono-, di- and tri-nucleotide at every position in the probe (one if present at a position, zero otherwise; similar to the implicit representation in the WD kernel). The resulting feature vector was augmented with the sequence derived features to form $\boldsymbol{x}_i$. In training, the λ-regularized quadratic error is minimized [70]:

$$\min \lambda ||w||^2 + \sum_{i=1}^{n} (\mathbf{w}^T \mathbf{x}_i - \hat{y}_i)^2$$

with

$$\mathbf{w} = \left(\lambda I + \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \right)^{-1} \sum_{i=1}^{n} \hat{y}_i \mathbf{x}_i$$

being its solution. Then $f_q(\boldsymbol{x}) = \mathbf{w}_q^T \boldsymbol{x}$ was the resulting regression estimate.

Ridge regression was straightforwardly implemented relying on Matlab's efficient matrix operations. As it generated virtually identical results to SVR, but is much less demanding in terms of computation time, it became a constitutive component of our tiling array normalization pipeline.

[3] `http://www.shogun-toolbox.org`

2.3.4 Tiling Probe Annotation

Because it was required for several analyses we generated an annotation of each individual tiling probe specifying whether its sequence was included in an annotated exon or intron (and to which gene that belongs) or whether the corresponding genomic region is annnotated as intergenic. We therefore mapped tiling probes to *Arabidopsis* gene models using the TAIR7 annotation [178]. We labeled as exonic, intronic or intergenic only those probes that were in their entire length included in an annotated segment. All other probes were labeled as ambiguous and ignored for the purpose of training and performance evaluation of transcript normalization or transcript mapping. Ambiguous probes included ones that span a transcript end or splice junction or ones complementary to regions where different gene models overlap or where annotated transcript isoforms of the same gene differ in their exon-intron structure.

One the one hand, the resulting tiling probe annotation was useful as a label sequence for transcript mapping, i.e., the "true" sequence of expression states needed for training an evaluation of such a method (see Section 2.6). On the other hand, the tiling probe annotation allowed us to define a tiling probe set for each annotated gene as follows: From all probes mapped to exons (either coding or untranslated region) in their entire length we retained only those for expression analysis which correspond to constitutive exons in all annotated splice forms of the same gene. We further excluded probes mapped to more than one (overlapping) gene model, and in order to reduce cross-hybridization artifacts, we also removed repetitive probes whose 25mer sequence occurred multiple times in the genome (see Section 2.1.2). For expression measurements from tiling arrays, we only considered the set of 30,228 annotated genes that are represented by at least three probes. This tiling probe set definition was the basis for profiling the expression of nearly 10,000 genes not represented on the ATH1 array and thus missed in previous expression analyses with GeneChip© microarrays. Furthermore, introns and exons can be monitored individually — a prerequisite for the detection and quantification of alternative transcript isoforms. The results of tiling array-based expression analyses were presented in Laubinger et al. [96, 97], Naouar et al. [123], Zeller et al. [205].

2.4 Label Sequence Learning with Hidden Markov Support Vector Machines

In label sequence learning, one is given a sequence of observations, which can for instance be a sequence of hybridization intensities from a tiling path, i.e., tiling probes with consecutive genome coordinates. Then the task is to assign a label to each observation, for example to distinguish between probes on a tiling array which interrogate exons from probes in nonexpressed regions.

In computational biology, Hidden Markov Models (HMMs) are very popular and have been the method of choice for solving label sequence learning problems [44]. Recently, a new inference method, namely HM-SVMs, has been developed in the field of machine learning [3, 137, 184]. It can be seen as an extension of Support Vector Machines (SVMs). HM-SVMs have been successfully applied in natural language processing [e.g., 3, 159], computational gene finding [140], and spliced sequence alignment [12, 154]. This diversity illustrates the flexibility and power of the approach.

HMMs and HM-SVMs take essentially the same modeling approach, but differ in the way parameters are estimated in training. While HM-SVMs use discriminative large-margin techniques related to SVMs, HMMs are generative models that attempt to estimate probability densities over the observation sequence and the corresponding segmentation. However, it has been argued that generative approaches do not lead to the best discrimination performance, as high-dimensional density estimation is known to be a harder task than discrimination [125, 126, 186]. One reason generative methods are often outperformed by discriminative methods is that they typically need to assume independence between observations in a sequence and also between features if several are used for learning. Since the HM-SVM method does not assume independence, it is very well-suited for many tasks in genome research for which measurements are dependent.

In addition to HM-SVMs, other discriminative structured output prediction algorithms, most prominently Conditional Random Fields (CRFs) [94], have successfully been applied to bioinformatics problems, such as gene finding or RNA secondary structure prediction [37, 40, 63].

In the following sections, I will review the HM-SVM learning algorithm, describe a powerful explicit feature map (an extended linear kernel) and finally empirically assess and discuss basic properties of HM-SVMs in comparison to HMMs.

2.4.1 Label Sequence Learning Problem

Formally, to solve the label sequence learning problem, we would like to learn a function

$$f : X \to \mathcal{S}^\star$$

that predicts a label sequence (more precisely a sequence of states, or simply a path) $\pi \in \mathcal{S}^\star$ given the sequence of observations $\mathbf{x} \in X$ (input features), both of equal length t, where $\mathcal{S}^\star$ denotes the Kleene closure. This is done indirectly via a discriminant function

$$F : X \times \mathcal{S}^\star \to \mathbb{R}$$

that assigns a real-valued score to a pair of observation and state sequence [3]. Once F is known, f can be obtained as

$$f(\mathbf{x}) = \underset{\boldsymbol{\pi} \in \mathcal{S}^\star}{\operatorname{argmax}} F(\mathbf{x}, \boldsymbol{\pi}).$$

In our case F satisfies the Markov property, and, consequently, this decoding can be computed efficiently by dynamic programming using the Viterbi algorithm [44, 57].

2.4.2 State Model

Allowed transitions between states are conveniently specified by a graphical state transition model. In such a graphical model, states are represented by nodes and allowed transitions are represented by arcs. Most of our knowledge about the problem is encoded in the structure of this state model. Examples are given below for the problems of polymorphic region prediction from resequencing arrays (Fig. 2.4) and transcript mapping from transcriptome tiling arrays (Fig. 2.5).

2.4.3 Parametrization

The goal of training an HM-SVM is to learn an optimal discriminant function $F_{\boldsymbol{\theta}}$ (parametrized by $\boldsymbol{\theta}$) which in conjunction with efficient decoding yields an optimal predictor of label sequences f.

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

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

where $[[.]]$ denotes the indicator function. For convenience of notation we assume a pseudo-transition $\phi(\pi_0, \pi_1) = 0$.

In this work, we modeled the feature scoring functions $g_{j,k}$ as piecewise linear functions as follows [similar to 140]: Let S be the number of supporting points s_l (satisfying $s_l < s_{l+1}$) and v_l their values, then the piecewise linear function is defined by

$$g(x) = \begin{cases} v_1 & x < s_1 \\ \frac{v_l(s_{l+1}-x)+v_{l+1}(x-s_l)}{x_{l+1}-x_l} & s_l \le x < s_{l+1} \\ v_S & x \ge x_S \end{cases}$$

Supporting points on the abscissa are typically chosen such that in each interval $[s_l, s_{l+1}]$ there are 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 $\theta_{j,k,l}$ constitute the parametrization of the model (denoted by $\boldsymbol{\theta}$).

2.4.4 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 want 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.,

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

To achieve this, the following optimization problem is solved:

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

$$\text{s.t.} \qquad F_\theta(\mathbf{x}^{(i)}, \boldsymbol{\pi}^{(i)}) - F_\theta(\mathbf{x}^{(i)}, \overline{\boldsymbol{\pi}}) \geq 1 - \xi^{(i)} \quad \forall \overline{\boldsymbol{\pi}} \neq \boldsymbol{\pi}^{(i)} \quad \forall i = 1, \ldots, n \tag{2.1}$$

where Ω is an additional regularization term. Regularization is a technique commonly used in empirical inference to avoid overfitting and to improve generaliztion. Regularization strength can be adjusted using the hyper-parameter C.

A soft-margin is implemented by so-called slack variables $\xi^{(i)}$ [32] allowing some prediction errors on the training set.

Except for the regularizer Ω, which will be discussed in more detail later, $F_{\boldsymbol{\theta}}$ is linear in all parameters and hence the constraints in (2.1) are linear. In case of a linear regularizer, we thus have to solve a linear programming problem (LP), whereas a quadratic regularizer leads to a quadratic programming problem (QP).

Because there are exponentially many wrong paths $\overline{\boldsymbol{\pi}}$, we also have an exponential number of margin constraints in (2.1), which prohibits solving the optimization problem directly. Instead, starting from an empty set of margin constraints and an arbitrary parametrization $\boldsymbol{\theta}^{(1)}$, we maintain an increasing working set of constraints corresponding to paths which maximally violate the margin. For this we use the Viterbi algorithm [44] which decodes the best path for a given parametrization and loss function. If it returns the true path, no margin violation occurred for this example; otherwise we generate a new constraint from the true path (known in the training set) and the wrong path (returned by Viterbi decoding). Adopting a column generation technique, adding constraints and solving the intermediate LP / QP is alternated till convergence to the (provably) optimal solution [3, 67, 138] (Table 2.1). The intermediate LPs or QPs are solved using either the CPLEX[4] or the Mosek[5] optimization software, both of which facilitate training with several thousand example sequences. In practice, we commonly resort to two heuristics to reduce the computational cost of HM-SVM training: First, before solving intermediate training problems, we do not consider constraints that were inactive by more than a certain margin in the last iteration. The size of this margin depends on the loss (see ($\star$) in Table 2.1). Second, we terminate the training when the objective function has not changed substantially during the last three iterations (less than 10^{-3} of its current value, see ($\star\star$) in Table 2.1). Both of these heuristics only marginally affect the accuracy of the learned model, while they speed up training significantly.

```
start at $t = 0$ with an arbitrary parametrization $\boldsymbol{\theta}^{(t)}$
        and an empty working set of constraints $W = \emptyset$
do
        for each training example $(\mathbf{x}^{(i)}, \boldsymbol{\pi}^{(i)})$
                decode the maximal margin violator $\overline{\boldsymbol{\pi}}$ using the Viterbi algorithm
                add a constraint of the form
                        $F_{\boldsymbol{\theta}^{(t)}}(\mathbf{x}^{(i)}, \boldsymbol{\pi}^{(i)}) - \max_{\overline{\boldsymbol{\pi}} \neq \boldsymbol{\pi}^{(i)}} \{F_{\boldsymbol{\theta}^{(t)}}(\mathbf{x}^{(i)}, \overline{\boldsymbol{\pi}})\} \;\geq\; 1 - \xi^{(i)}$
                        to the working set $W$
        end for
        solve the intermediate training problem with the updated working set $W$
                to obtain the next parametrization $\boldsymbol{\theta}^{(t+1)}$                ($\star$)
        $t = t + 1$
until no more constraints were added to $W$                                                      ($\star\star$)
```

Table 2.1: The HM-SVM training algorithm in each iteration alternates between constraint generation and QP/LP solving.

2.4.5 Linear and Quadratic Regularization

Depending on the application, we used linear or quadratic regularizers. For each of them, we give an example below.

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

This regularizer implements the idea that absolute parameter values should be small and with the second term, large changes between the values v_l and v_{l+1} at adjacent supporting points of the same feature scoring functions are penalized. This amounts to penalizing large variation of the feature scoring functions (with respect to the choice of supporting points) to obtain “smoother” or “simpler” functions, which appear more parsimonious.

[4] www.cplex.com

[5] www.mosek.com

The analoguous quadratic regularizer has the form

$$\Omega(\boldsymbol{\theta}) = \boldsymbol{\theta}\ ^{2} + \sum_{j=1}^{m} \sum_{k \in \mathcal{S}} \sum_{l=1}^{S-1} (\theta_{j,k,l} - \theta_{j,k,l+1})^2$$

Quadratically penalizing the variation typically results in more gradually changing feature scoring functions as compared to ones that are linearly regularized and often appear more like step functions.

Furthermore, additional properties of the feature scoring functions can be encoded in the regularizer using similar techniques. We can, e.g., couple scoring functions of different states k and k' via regularization on $(\theta_{j,k,l} - \theta_{j,k',l})^2$ or enforce monotonicity by additional constraints of the form $\theta_{j,k,l} - \theta_{j,k,l+1} \leq 0$ (for monotonic increase; alternatively ≥ 0 for monotonic decrease) for all l of some features j and states k.

2.4.6 Loss Function

In practice, the basic algorithm described above is augmented with a loss function Δ that encodes a problem-specific dissimilarity measure for a pair of label sequences. It allows us to adjust the loss, a predicted 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 that. We use the loss function to rescale the margin (although slack-rescaling has also been proposed) [3, 180], replacing the margin constraints in (2.1) with:

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

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

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

The loss function is required to be non-negative and decomposable for efficient decoding via dynamic programming. For instance, the Hamming loss is a simple function with the desired properties. Below, we will also show problem-specific loss functions encoding our prior knowledge about the problem.

2.4.7 Generative HMM Training with Identical Parametrization

In order to compare discriminative and generative learning algorithms, we used generative maximum likelihood training techniques to fit a model with identical parametrization. Transition probabilities were obtained as

$$\phi(k, l) = \frac{A_{k,l}}{\sum\limits_{l'} A_{k,l'}}$$

for all state pairs $(k, l) \in \mathcal{S}^2$. $A_{k,l}$ were counts of transitions observed in the label sequences π^i of the training set $(i = 1, \ldots, n)$

$$A_{k,l} = \sum_{i=1}^{n} \sum_{p=1}^{|\pi^i|} [[\pi^i_p = k \wedge \pi^i_{p+1} = l]]$$

where $|\pi^i|$ denotes the sequence length and $[[]]$ the indicator function [44].

For the estimation of emission probabilities we modified the standard HMM maximum likelihood estimation [44] as follows to account for piece-wise linear feature scoring functions $g_{j,k}$. The values of the supporting points v_l $(l = 1, \ldots, S)$ of $g_{j,k}$ for feature j and state k were estimated as

$$v_l = \frac{E_l}{\sum\limits_{l'} E_{l'}}$$

Here, E_l are contributions from feature values x around the supporting point s_l where k is the true state. With

$$c_l(x) = \begin{cases} 0 & x \le s_{l-1} \quad \vee \quad x \ge s_{l+1} \\ 1 & l = 1 \quad \wedge \quad x \le s_1 \\ \frac{x - s_{l-1}}{s_l - s_{l-1}} & 1 < l < S \quad \wedge \quad s_{l-1} < x \le s_l \\ \frac{s_{l+1} - x}{s_{l+1} - s_l} & 1 < l < S \quad \wedge \quad s_l < x < s_{l+1} \\ 1 & l = S \quad \wedge \quad x \ge s_S \end{cases}$$

E_l was obtained by summing up these position-wise weights over the training sequences $\mathbf{x}_j$ of the jth feature,

$$E_l = \sum_{i=1}^{n} \sum_{p=1}^{|\pi^i|} [[\pi^i_p = k]]\, c_l(x^i_{j,p})$$

Finally, probabilistic parameters were log-transformed to be able to use the same additive Viterbi decoding algorithm as for the HM-SVM.

2.4.8 Properties of HM-SVMs in Comparison to HMMs

We explore some of the properties of HM-SVMs, starting with an empirical analysis of their run-time and generalization accuracy. For comparison, we also evaluated an identically parametrized model fitted with the maximum-likelihood HMM training method. All experiments in this section were based on the mSTAD model as an example (see Section 2.6) using the D_001 data set (see Table 4.4 for details). mSTAD was implemented in Matlab with performance-critical components written in C++. The mSTAD implementation has 621 parameters and additionally 1,528 auxiliary variables (for the regularizer). Together with the slack variables (one per training example) the quadratic training problem can have more than 4,000 free variables.

To facilitate a meaningful comparison, the examples for HMM training had to be selected more carefully than for the HM-SVM. Here, we chose the selection strategy that was found to be most favorable for mSTAD HMM (see Section 2.6.5 for details). Therefore, the fact that training sets differed, poses a caveat to the interpretation of the results. However, this can only be an advantage for HMMs in the following comparisons.

As a result of our first experiment, we found significantly higher (asymptotic) test accuracy for the HM-SVM than for the HMM. Moreover, HM-SVMs seemed to approach their asymptotic error faster (w.r.t. training set size) than HMMs (Table 2.2).

	Training set size							
	15	30	60	125	250	500	1,000	2,000
Total length	1.4K-2.3K	3.1K-3.6K	6.2K-6.3K	12K-14K	25K-26K	50K-52K	100K-104K	201K-209K
Training time								
HM-SVM (QP)	4-5m	4-6m	6-8m	8-13m	19-30m	92-66m	3-4h	17-20h
HMM	0.6s	1.4-1.6s	2.3-2.4s	4.8-5.1s	9.3-10.2s	20-25s	47-48s	81-82s
Test accuracy [%]								
HM-SVM (QP)	78.2-81.2	79.2-81.5	80.1-81.5	80.4-81.7	79.9-80.8	80.2-81.5	80.3-81.5	80.8-82.0
HMM	68.9-73.7	69.4-78.4	71.9-78.2	74.3-77.5	75.1-77.0	75.8-78.1	75.7-77.3	76.0-77.8

Table 2.2: Generalization accuracy and training time of HM-SVMs in comparison to HMMs. We used the CPLEX optimization software (version 9) for solving the (intermediate) quadratic training problem(s). All values are given with respect to three different training and test sets (3 cross-validation folds with the same number of sequences). Note that small inaccuracies in timing can also be due to the fact that experiments were carried out on a heterogeneous compute cluster. As test error, we report the average of precision and recall for exon probe recognition (as percentages). The same regularization strength has been applied throughout, although this might lead to suboptimal results for some training set sizes (see below for a model selection).

Next, we explored the robustness of both, the HMM and the HM-SVM, with respect to label noise. We determined their performance on three cross-validation folds (each containing 500 training sequences and 500 test sequences) to account for random fluctuations. Label noise was introduced by randomly choosing a certain proportion of segments and converting their label to another type (from exons to introns and vice versa where the result was still a valid gene segmentation, see also Section 2.6 for details of mSTAD's learning task). Both label sequence learning methods were found to be relatively robust to low noise levels ($< 10\%$), but for higher levels, HMM performance declined more dramatically than that of HM-SVMs (Table 2.3). This is consistent with the finding that the training set for HMMs had to be pre-filtered, which can be seen as removing unreliable, more noisy example sequences (see also Section 2.6.5).

In another series of experiments, we evaluated the robustness of the HM-SVM learning algorithm with respect to the hyperparameter C, which controls regularization strength, i.e., the balance between fitting the training data and model complexity. These experiments were carried out on fixed training and test samples with 500 examples each. We observed that, generally, training time increased with decreasing regularization strength. One explanation for this is that there is relatively more freedom to fit a weakly regularized model and, hence, more constraints have to be generated until convergence. Too strong regularization on the other hand resulted in suboptimal performance (Table 2.4). We also conducted a model selection for HMM training with varying pseudocounts. However, here

	Proportion of (segment) label noise						
	0%	1%	2%	5%	10%	20%	50%
Test accuracy [%]							
HM-SVM (QP)	80.2-81.6	80.3-81.4	80.1-81.3	80.2-80.9	79.7-80.5	79.4-79.8	78.2-78.8
HMM	75.8-78.1	74.0-76.2	73.6-76.1	71.0-75.4	69.9-72.0	66.1-69.6	59.1-67.5

Table 2.3: Generalization accuracy of HM-SVMs and HMMs as a function of artificially introduced label noise. The range of accuracies was determined on three cross-validation folds (with 500 training and test sequences each) and resulted from evaluating exon probe recognition in percent. The percentage of noise is given relative to the number of (exon and intron) segments which can be converted without generating invalid (gene) segmentations; the proportion of wrong labels among all labels is substantially lower. Also note that, due to differences in the training sets for the HMM and the HM-SVM, the exact number of converted segments also differs.

we found that Laplace' rule (adding 1 to each unused parameter weight) worked best and larger pseudocounts led to worse results (data not shown).

	Regularization strength C					
	500	50	5	0.5	0.05	0.005
Training time [m]						
HM-SVM (QP)	49	23	34	69	87	192
Test accuracy [%]						
HM-SVM (QP)	76.5	78.7	79.6	80.2	80.1	80.2

Table 2.4: Generalization accuracy and training time of HM-SVMs as a function of regularization strength, which is controlled with the hyperparameter C. All values are given for a single training and test set of 500 examples each. As before, we used exon probe recognition in percent as our evaluation metric. (mSTAD actually employs a more complicated regularizer, but here we systematically varied only C_2, for simplicity denoted by C. C_1 was always set to $0.02 \times C$ and C_3 to $0.2 \times C$, see Section 2.6.4 for more details.)

Finally, we empirically assessed the influence of the choice of the loss function: First, we considered one which is adapted to the problem at hand; second, we used the generic Hamming loss. With the same implementation and experimental setup as before, we simply replaced mSTAD's loss function (see Section 2.6.4) with the Hamming loss and reported the generalization accuracy for a range of values for the hyperparameter C (Table 2.5). Another model selection was needed here, because the Hamming loss (theoretically) scales in a manner that is different from mSTAD's original loss (see also Section 2.4.4). For the Hamming loss, the best model was found for the smallest value of C. However, fitting models with weaker regularization turned out to be impractical due to convergence issues. A comparison of the best models for the respective methods revealed that, indeed, the problem-specific loss function contributes substantially to the HM-SVM's model accuracy (Table 2.5).

	Regularization strength C					
	500	50	5	0.5	0.05	0.005
HM-SVM test accuracy [%]						
using mSTAD's loss	76.5	78.7	79.6	80.2	80.1	80.2
using Hamming loss	76.7	78.0	78.0	78.1	78.1	78.2

Table 2.5: Generalization accuracy of HM-SVMs using a problem-specific loss (as in Table 2.4) or the Hamming loss (see Section 2.6.4 for details of mSTAD's regularizer and loss).

2.4.9 Discussion of HM-SVM Properties

In this work we augmented the HM-SVM learning algorithm proposed in Altun et al. [3] with an expressive explicit feature map utilizing piece-wise linear functions (as similarly proposed in Rätsch et al. [140]) [see also 180, 184]. While the details of its application to problems arising in genome biology are given in the following sections, here we explored some basic properties of HM-SVMs in comparison to another label sequence learning algorithm, namely generative HMMs. Specifically, we found that higher label-wise accuracy can be achieved with HM-SVMs (for the application of transcript identification, see Section 2.6). However, this is associated with substantially larger training efforts: on data sets where HMM training takes less than a minute, training an HM-SVM can require several CPU hours, even when commercial high-performance optimization software is employed (Table 2.2). The fact that we rely on such software is an additional disadvantage for the distribution and free use of our HM-SVM implementations. Furthermore, achieving optimal performance with HM-SVMs requires to adjust regularization strength C via model selection. Our experiments, however, indicate that, in practice, a coarse grid search might be sufficient, since we noticed only marginal deviations from optimal test accuracy for values of C ranging across three orders of magnitude (Table 2.4). Another factor contributing to the high accuracy of HM-SVMs is a reasonably calibrated loss function: with the simple Hamming loss, HM-SVM accuracy (78.2%) is similar to that of HMMs (75.8-78.1%, see Tables 2.2, 2.5). While our experiments indicate that carefully designing the HM-SVM loss function pays off in terms of accuracy, this certainly poses an additional challenge for successfully applying HM-SVMs to real-world problems — much like the choice of an appropriate kernel for the problem at hand is crucial for obtaining an SVM classifier with optimal performance. Despite all the limitations of HM-SVMs discussed here, it is of note that already with a small training set, containing < 50 sequences, a better generalization error is obtained with HM-SVMs than HMMs could achieve on any training set (up to 2,000 examples). For such a small data set, HM-SVM training is very practical, requiring only 5 minutes with our implementation. In addition, our empirical results indicate that HM-SVMs are much more tolerant to label noise than HMMs. This property is highly desirable for any learning algorithm applied to biological data, because these are very rarely free of label noise.

When comparing different label sequence learning methods on prediction tasks very dissimilar to ours, Nguyen and Guo [126] still arrived at the same conclusion that HM-SVMs performed better than HMMs. On these problems HM-SVMs were found to perform even better than other discriminative label sequence learning methods such as CRFs [94]. For classification problems, it has long been a consensus that discriminative methods outperform their generative counterparts given that there is sufficient training data, i.e., that discriminative classifiers have lower asymptotic error. An intuitive explanation for this is that generative models, which learn the joint probability $P(x, y)$ of inputs x and labels y, try to solve a more general and, hence, more difficult problem than discriminative ones directly modeling $P(y|x)$, which is sufficient for classification [125, and references therein]. Ng and Jordan [125] provide theoretical and empirical evidence for this in a rigorous comparison of a Generative-Discriminative pair of classifiers.

On a more general level, another important difference between HMMs and HM-SVMs can be added: while the latter are limited by their supervised learning approach, HMMs can be trained in an unsupervised fashion [44] or using both supervised and unsupervised techniques [e.g., 120]. However, when there are sufficiently many labeled examples — even of dubious quality — HM-SVM training may yield more accurate predictions than HMM maximum likelihood training.

2.5 Engineering HM-SVMs for the Detection of Polymorphic Regions

This section describes a first application of HM-SVMs to the detection of regions of elevated polymorphism levels from resequencing array data and is based on joint work with Richard M. Clark, Korbinian Schneeberger, Anja Bohlen, Detlef Weigel and Gunnar Rätsch (see p. 137 for author contributions) [203]. The resulting method is called margin-based prediction of polymorphic regions (mPPR). It specifically addresses the extraction of label data from experimentally characterized polymorphisms, the generation of relevant features as well as adaptations of the HM-SVM algorithm to the problem at hand. Furthermore, details of performance evaluation and biological validation are given. In the following, I present joint work with Richard M. Clark, Korbinian Schneeberger, Anja Bohlen, Detlef Weigel and Gunnar Rätsch (see p. 137 for author contributions) [203].

2.5.1 Preparation of Hybridization, Repeat and Sequence Data

Our predictor of polymorphic regions was specifically developed for the analysis of a previously published *Arabidopsis* resequencing array data set, hereafter called “AtAD20” [29]. It contains data generated for $> 99.99\%$ of bases in the 119 Mb reference genome [73] for 20 accession including the reference Col-0 [29]. For training an evaluation a made use of a previously published polymorphism resource in *Arabidopsis* [128]. It contained 1,213 fragments of ≈ 550 bp in length, which had been sampled by PCR and dideoxy sequencing throughout the genome for 19 of the 20 AtAD20 accessions. This data set, hereafter called “2010”, covers about 0.5% of the genome per accession, and comprises ≈ 2700 SNPs and ≈ 400 indel polymorphisms per target accession [128].

Quantile normalized hybridization data from Clark et al. [29] facilitated the use of predictors trained with data from all accessions. Consequently, predictors were available to make predictions on any accession. We used a data set (2010) of previously characterized polymorphisms to generate the label set for both PRs and conserved regions (see Section 3.1.1; [29, 128]). Array measurements for repetitive oligonucleotides are much less reliable than for unique oligonucleotides; therefore, we annotated repetitive 25mer oligonucleotides on the resequencing arrays as described before (see Section 2.1). We combined information for all types of 25mer repeats 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.

2.5.2 Overview of the mPPR Algorithm

The graphical model underlying mPPR is displayed in Fig. 2.4. 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 HM-SVMs [3] for label sequence learning. We adapted these by defining an appropriate loss function, detailed below (Table 2.7). From the predicted state sequence we afterwards inferred the label sequence (see color coding in Fig. 2.4).

2.5.3 State Model

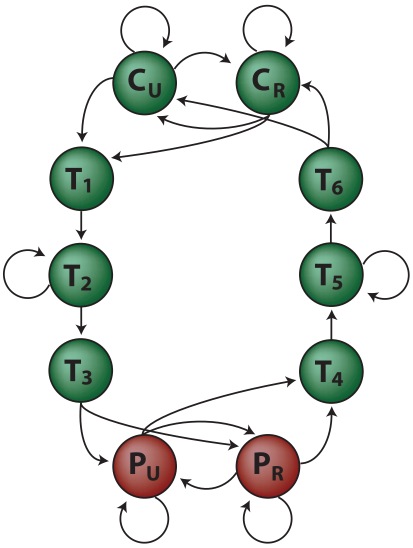


Figure 2.4: State model employed in mPPR. States are drawn as colored circles, transitions as arrows. $\mathtt{C}_\mathtt{U}$ and $\mathtt{C}_\mathtt{R}$ model conserved sites (unique and repetitive, respectively). Similarly, $\mathtt{P}_\mathtt{U}$ and $\mathtt{P}_\mathtt{R}$ model sites that are polymorphic 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 (see also Fig. 3.1).

The simplest possible model, with one state C for conserved nucleotides and one state P for polymorphic regions was extended in two ways. First, we noted that hybridization signal gradually decreases over a few nucleotides towards a polymorphism. We therefore included a series of three states $\mathtt{T}_1, \mathtt{T}_2, \mathtt{T}_3$ modeling decreasing intensities upstream of polymorphic regions and similarly three states $\mathtt{T_4}, \mathtt{T_5}, \mathtt{T_6}$ for increasing intensities downstream (Fig. 2.4 for details). The second extension relates to repetitive sequences, which we modeled separately from unique sequences via duplicated states which effectively allowed feature scoring functions for repetitive regions to be learned differently. The model contains a state $\mathtt{C_R}$ for conserved, repetitive sequences (positions p where $RM(p) = 1$) and a state $\mathtt{C_U}$ for conserved, unique sequences (where $RM(p) = 0$); likewise, a state $\mathtt{P_U}$ for polymorphic, unique sequences with $\mathtt{P_R}$ as the repetitive counterpart. Transition states $\mathtt{T_i}$ were not duplicated. We denote the set of states by $\mathcal{S}$. Allowed transitions between the states are drawn as arcs in Fig. 2.4. Real-valued scores $\phi(i, j)$ were associated with transitions from state $i \in \mathcal{S}$ to state $j \in \mathcal{S}$, 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 $\phi(i, P_U)$).

2.5.4 Generation of Labelings

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 $\mathtt{P_U}$ or $\mathtt{P_R}$ states depending on the repeat annotation. In the next step, we assigned $\mathtt{P_U}$ or $\mathtt{P_R}$ states to sites between two polymorphic labels at a distance of ≤ 18 bp (for the choice of this distance see Fig. 4.2). Every segment of $\mathtt{P}$ states was then extended 6 bp in each direction and the transition states $\mathtt{T_1}, \ldots, \mathtt{T_3}$ and $\mathtt{T_4}, \ldots, \mathtt{T_6}$ were inserted upstream and downstream of every segment of $\mathtt{P}$ states, respectively. Finally, $\mathtt{C_U}$ or $\mathtt{C_R}$ states were assigned to the remaining positions. This procedure generated a state sequence for every fragment in the 2010 data set.

2.5.5 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, features directly derived from array intensities (Table 2.6, 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 [29]). 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 $\mathtt{C_U}$ and $\mathtt{C_R}$ states, as well as between $\mathtt{P_U}$ to $\mathtt{P_R}$, and not for learning *per se*.

Feature	Feature value derivation
1	$IM_t(p) = \frac{1}{2}\left[\log(I^+_{max}(p)) + \log(I^-_{max}(p))\right]$
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 2.6: Features used for polymorphic region prediction. Here we use the notation from the supplement of Clark et al. [29]: In general, superscripts + or − denote the strand, and subscripts Col or t denote the reference and target accession respectively (in the following collectively referred to as acc), 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 p' around p where $B^s(p') = seq(p') \quad \forall s \in \{+,-\}$. (For further details see supplement of Clark et al. [29].) All features were standardized prior to training (mean and standard deviation were estimated on the training set).

2.5.6 Problem-Specific Regularization and Loss

For the prediction of PRs we employed an HM-SVM algorithm as detailed above (see Section 2.4) using the linear regularizer introduced in equation 2.4.5. We further chose a position-wise loss function $\ell(p)$, which is summed over the whole sequence (of length t): $\Delta = \sum_{p=1}^{t} \ell(p)$ (similar to a weighted Hamming loss). Details are given in Table 2.7.

2.5.7 Cross-Validation, Evaluation and Whole-Genome Predictions

For 5-fold 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 [29]). Furthermore, known PRs as well as predicted PRs were excluded from precision-recall estimation if they contained

Predicted 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 2.7: Position-wise loss $\ell(p)$. We used a "tolerance" region, comprising non-polymorphic nucleotides in labeled blocks (up to 9 bp upstream and downstream of polymorphisms), where neither $\mathtt{C}$ nor $\mathtt{P}$ 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.

$\geq 75\%$ repetitive sites.

Replacing transition scores $\phi(i, i)$, $i \in \{\mathtt{C}_\mathtt{U}, \mathtt{C}_\mathtt{R}\}$ after training by $\hat{\phi}(i, i) = \phi(i, i) + \delta$ resulted in predictions either with increased precision ($\delta > 0$) or with increased recall ($\delta < 0$). Fifty-one values for δ were uniformly chosen from the interval $[-3, 2]$ to generate precision-recall curves for all 5 test subsets. For Fig. 3.3 and Fig. 4.3, precision-recall curves were averaged over the subsets.

The sequence type of each nucleotide was determined based on the TAIR6 *A. thaliana* genome annotation [178]. 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.

Precision and recall for whole-genome predictions are expected to be slightly different from the values estimated on the 2010 set as coding sequences are relatively overrepresented in 2010 compared to the entire genome [29, 128]. 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 and n^G_{cod} the number of coding bases in the genome. Then, for the whole genome, the number of true positives in coding regions is estimated as $TP^G_{cod} = \frac{n^G_{cod}}{n^{MN}_{cod}} TP^T_{cod}$. Applying the same corrections for false positives (FP), true discoveries (TD) and false negatives (FN), as well as for intergenic (ige) and UTR / intron bases (utr), precision 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 recall 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 precision, transition scores were independently tuned by choosing the smallest δ for which each of the predictors achieved precision $\geq 90\%$ 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 $\mathtt{C}_\mathtt{U}$ or $\mathtt{C}_\mathtt{R}$). 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.

2.5.8 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 [76] available at TAIR.[6] Only contigs of length ≥ 1 kb and containing only called nucleotides (i.e., A,C,G,T) were included in subsequent analyses. Using BLAT [88], 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; [76]), 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 precision and recall 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) [163, 179] spanning the S-locus region to the reference genome sequence with the alignment program stretcher in the EMBOSS package [119, 142]. 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 precision and recall as detailed above. To correct for the large disagreement between the two sets of PR labels (Monsanto and 2010, Table 3.5), which is most likely the result of sequencing or assembly errors in the shotgun Monsanto data, we multiplied the corresponding recall estimate by the fold difference in recall estimates for predictions evaluated on 2010 and on the genomic sequences for which the data sets overlapped (a factor of about 1.5; see Table 3.5).

[6] `http://www.arabidopsis.org/Cereon`

2.5.9 Ability of mPPR to Predict Long Deletions

We assessed the ability of our method to detect long deletions, which were absent in 2010, our training data, by using a test set of known deletions in the AtAD20 accessions [29]. 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 (Table 4.3). 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 (Table 4.3 and [29]). 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. [29] for a discussion). Minimally, however, our method was highly effective at identifying the approximate locations of long deletion polymorphisms in unique sequences (Fig. 3.6 B and Fig. 3.7).

2.5.10 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 [61, 160]. Genomic DNA was prepared from three week old seedlings with standard methods. For PCR, primers flanking *RPM1* were design using Primer 3.0 [146]; the predictions themselves were used to select primer pairs likely to hybridize to target sequences without mismatches (see supplement of Zeller et al. [203] for primer sequences and details of PCR and sequencing protocols). Sequence reads for each accession were aligned to the reference genome sequence using the program MUSCLE [46] with a gap open penalty of 1000 and a gap extension penalty of 10^{-6}. Alignments were then refined manually and converted into a graphical representation (Fig. 3.7).

2.5.11 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. 3.8 and Fig. 4.5). 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 [29].

2.5.12 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 [178]. 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 SNP data from MBML2. 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 transcript isoform.

2.5.13 Relation of PRs to Predicted Cis-Elements

Position-wise *cis*-element density was calculated using the predictions of O'Connor et al. [129] that were based on putative binding sites for 105 transcription factors (TFs). That the overlap between PRs and *cis*-elements is highly unlikely to be a random observation was established by permutation tests [supplement of 203].

2.5.14 Annotation of Predictions Relative to Genes

We calculated the overlap of PRs to coding sequences based on the TAIR6 annotation [178] with gene family descriptions as previously reported [29]. When mapping PR predictions to miRNA genes, we used the following divisions: precursor end (miRNA arm), miRNA, loop region, miRNA*, precursor end (miRNA* arm) (Fig. 3.12). Since the location of the miRNA* is not annotated in RFam [62], we calculated a secondary structure for each miRNA using RNAfold [71]. 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 $\alpha = l_r/m_r$ (Fig. 3.12).

2.6 Engineering HM-SVMs for Transcript Identification from Tiling Array Data

In this section we describe mSTAD (**m**argin-based **S**egmentation of **T**iling **A**rray **D**ata), an HM-SVM-based algorithm for transcriptional tiling array data. It incorporates ideas that were similarly presented before [56, 72], but here we used a different strategy for learning and inference — i.e., discriminative HM-SVM training rather than generative modeling (see Section 2.4). The mSTAD algorithm was developed together with with Stefan R. Henz, Sascha Laubinger, Detlef Weigel and Gunnar Rätsch (see also p. 137) and published in Zeller et al. [204]. Evaluations and biological applications are based on Laubinger et al. [97], Zeller et al. [205] and are the result of a collaboration with Sascha Laubinger, Stefan R. Henz, Timo Sachsenberg, Christian K. Widmer, Naira Naouar, Marnik Vuylsteke, Bernhard Schölkopf, Gunnar Rätsch and Detlef Weigel (see p. 137 for author contributions).

2.6.1 Problem Description and Modeling Approach

Our goal was to characterize each probe in a tiling path as either intergenic (not transcribed) or as part of a transcriptional unit (either exon or intron). Instead of predicting the label (intergenic, exonic or intronic) directly, mSTAD was trained to associate a state with each probe given its hybridization measurements and the local context. From the state sequence one can easily infer the label sequence (see color coding in Fig. 2.5). For learning, we first had to define the target state sequence, i.e., the "truth" that we tried to approximate. It was generated using known transcripts from the TAIR7 annotation [178] together with hybridization measurements. We then applied HM-SVMs [3] for label sequence learning to build a discriminative model capable of predicting the state and hence the label sequence given the hybridization measurements alone.

2.6.2 State Model

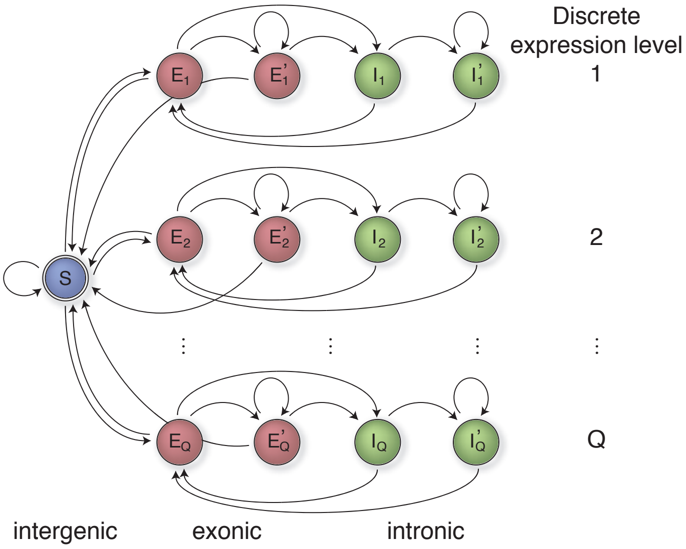


Figure 2.5: State model employed in mSTAD. For each of $Q = 20$ discrete expression levels there is a submodel consisting of two exon and two intron states. Modeling introns depending on the expression level of the surrounding exons allows carrying expression information along the whole transcript. Colors indicate the labels intergenic (blue), exonic (red) and intronic (green). Duplicated exon and intron states were found to result in slightly improved performance compared to a model with single exon and intron states.

A simple version of the state model had only three states: intergenic, exonic and intronic. We extended it in two ways: (i) by introducing an intron / exon start state that allows modeling of the start and the continuation of exons and introns separately and (ii) by repeating the exon and intron states for each expression quantile. This allowed us to model discrete expression levels separately (Fig. 2.5). To compensate for uneven intensity decreasing from the 3' transcript end (see Section 3.2.4), we additionally allowed transitions from the exon states of an expression quantile to the next higher or lower level.

2.6.3 Generation of Labelings

For genomic regions with known transcripts we considered the sense direction of up to 1 kb flanking intergenic regions while maintaining a distance of at least 100 bp to the next annotated gene. Within this region we used the probe annotation described in Section 2.3.4 as the "true" label sequence. In a second step we subdivided genes according to the median hybridization intensity of all exonic probes into one of $Q = 20$ expression quantiles. For each probe, a state was determined from its label and expression quantile. Probes with ambiguous annotation, i.e., ones spanning transcript ends or splice junctions and ones complementary to multiple genes or transcripts, were excluded from performance evaluations.

2.6.4 Problem-Specific Regularization and Loss

In mSTAD, we incorporated a quadratic regularizer of the form

$$\Omega(\boldsymbol{\theta}) = C_1 \ \boldsymbol{\theta}\ ^2 + C_2 \sum_{j=1}^{m} \sum_{k \in \mathcal{S}} \sum_{l=1}^{s-1} (\theta_{j,k,l} - \theta_{j,k,l+1})^2 + C_3 \sum_{j=1}^{m} \sum_{(k,k')} \sum_{l=1}^{S} (\theta_{j,k,l} - \theta_{j,k',l})^2$$

where (k, k') denotes either a pair of exon states or a pair of intron states with corresponding expression levels i and $i + 1$. In addition to penalizing the absolute parameter values and differences between parameters of adjacent supporting points within one feature scoring function, the third term constrained differences between feature scoring functions of exon (and intron) states of neighboring expression levels. In doing so, we encoded our preference for feature scoring functions which are similar between exon states (as well as for intron states). Feature scoring functions resulting from training on root tissue (D1, see Table 4.4) are shown in Fig. 2.6. Note that C_1, C_2 and C_3 allowed us to individually adjust regularization strength for each term. When training on 1000 regions containing one annotated gene each, model selection indicated optimal or nearly optimal generalization performance for $C_1 = 0.01$ (with $100 \times C_1$ for squared transition scores), $C_2 = 0.5$ and $C_3 = 0.1$ (data not shown). Interestingly, mSTAD's performance is relatively robust with respect to these hyperparameters: changes in prediction accuracy are insignificant even when hyperparameters change by an order of magnitude (see Section 2.4.8).

We designed a loss function $\Delta = \sum_{p=1}^{t} \ell(p)$, which is position-wise (probe-wise) decomposable over a sequence (a tiling path) of length t. The position-wise loss function $\ell(p)$ is detailed in Table 2.8. This loss function was designed to strongly penalize false positive exon predictions in intergenic regions. False negative exon and intron predictions incur only half the position-wise loss and a relatively low position-wise loss is incurred by exons for which the discrete expression level deviates from the "true" expression level in a manner increasing with level difference.

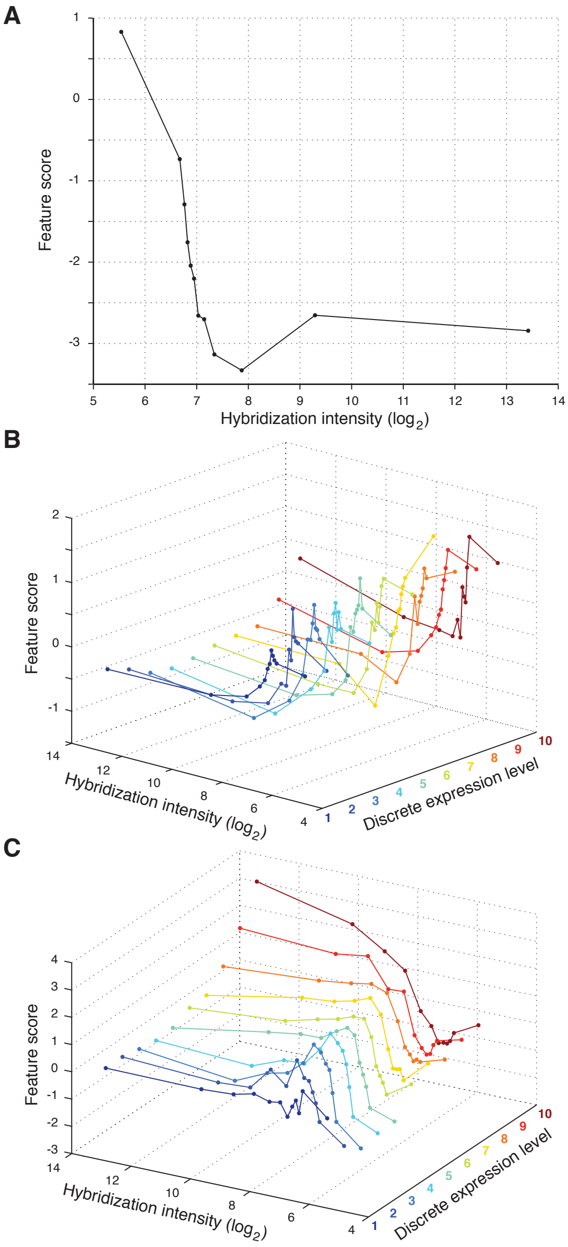


Figure 2.6: Feature scoring functions that mSTAD learned in training on tiling array data from root tissue (D1, see Table 4.4). **(A)** Feature scoring function, i.e., the transformation from hybridization intensity to feature score, utilized by the intergenic state (Fig. 2.4). **(B)** Feature scoring functions of intron states (I'_l, Fig. 2.6) by expression level (l, color-coded). **(C)** Feature scoring functions of exon states (E'_l, see Fig. 2.6) by expression level (l, color-coded).

2.6.5 Comparing mSTAD to Other Transcript Mapping Methods

For the comparison of HM-SVMs to HMMs and the transfrag method [85], predictions were generated as follows. Prediction sets with different trade-offs between precision and recall were generated by adjusting transition scores after mSTAD HM-SVM training. Manipulating the scores $\phi(k, k')$ associated with transitions leaving any exon state k, was found to yield a broad range of accuracy balances. Predictions were generated for $\phi(k, k') + \delta$ with $\delta \in \{-0.6, -0.5, \ldots, 0.5, 0.6\}$ (Fig. 3.19).

For mSTAD HMM (the same parametrization fitted with a generative HMM training algorithm, see Section 2.4.7), different balances between precision and recall were realized

Predicted state	Probe label		
	intergenic	intronic	exonic
S (intergenic)	0	0.5	0.5
I, I' (intron)	0.5	0	0.5
E, E' (exon)	1	0.5	0.1δ

Table 2.8: Position-wise loss $\ell(p)$. By δ we denote the difference between the predicted discrete expression level and the "true" expression level of an exon (inferred as the median intensity quantile of all exon probes interrogating the same annotated transcript). Except for this exon-exon loss term, the loss function is identical for all (predicted) expression levels. Regardless of the predicted state, no loss is incurred for probes with ambiguous label (e.g., probes spanning splice junctions or alternatively spliced portions of a gene). For details on the set of states, see Fig. 2.5.

by sampling training sequences around genes with different expression levels. For this, all annotated genes were partitioned into ten expression bins based on the median intensity of annotated exon probes. Subsequently eight different training sets were sampled either from all genes or the top $N\%$ expressed genes with $N \in \{30, 40, \ldots, 90\}$. As expected, precision of the HMM predictions increased with the minimum gene expression level in the training set used to fit the model.

Transfrags were computed as described by Kampa et al. [85] and implemented in the Affymetrix Tiling Analysis (TAS) Software version 1.1 build 2. We evaluated 900 different combinations of parameters on the same set of full-length cDNA-confirmed genes used for the assessment of mSTAD's performance. (bandwidth varied in steps of 25 between 50 and 150, signal threshold between 5 and 13, min run in steps of 20 between 20 and 100 and max gap in steps of 20 between 40 and 100).

For all methods prediction accuracy was determined on the same set of 1000 genomic regions each containing exactly one annotated gene that was in its entirety supported by full-length cDNA sequences. None of these regions overlapped with examples used for training or parameter tuning. For mSTAD HM-SVM and mSTAD HMM predictions were made according to a two-fold cross-validation scheme using 1000 disjoint examples for training; transfrags were directly evaluated on the test set. All evaluations shown (Fig. 3.19 and Fig. 3.21) are based on tiling array data from root tissue (D_001, see Table 4.4).

Precision and recall were assessed in comparison to annotated genes on the level of individual tiling probes, per exon, per intron and on the level of exon-boundaries (Fig. 3.19 for definitions).

2.6.6 Generating TARs for Developmental Data

After evaluating mSTAD's capabilities of accurately recognizing transcripts from tiling array hybridization data in comparison to existing methods, we turned to an application of mSTAD to data from *A. thaliana* tissues and developmental stages (see Table 4.4 for a list of samples analyzed).

Generating Genome-Wide TAR Predictions

After preprocessing the hybridization data using the above described pipeline (quantile normalization followed by transcript normalization, see Section 2.3.2 and Section 2.3.3) we applied the mSTAD algorithm (a version modeling $Q = 10$ discrete expression levels).

For each sample, we trained mSTAD separately on mean intensities across replicates and used the trained instance only for prediction of array data from the same sample. To obtain unbiased whole-genome predictions we employed cross-validation. After splitting the genome between pairs of neighboring genes, one instance of mSTAD was trained on 500 of these genic regions and hyper-parameters were tuned on another 500 genic regions. We trained and tuned a second instance of mSTAD on two further disjoint sets of 500 genes each. For region-wise whole-genome predictions, we chose the mSTAD instance that had not seen the particular region during training and hyperparameter tuning (or a random instance if neither of them had). From the predicted labeling of tiling probes we extracted exon segments by assigning the genomic coordinates corresponding to the start of the first and the end of the last probe of a run of consecutive exon labels. The resulting segmentations are available as gff-files and visualized in the At-TAX Generic Genome Browser [173].[7]

To determine overlap between TAR predictions and annotated regions, we used the TAIR7 annotation [178] and direct alignments with EST and cDNA sequences.[8] Sample-specific segments were obtained as residual after computing the overlap between predicted exon segments in the tissue of interest to those from all other tissues (Fig. 3.23 B). Similarly, we obtained predictions specifically made for polyA(+/-) conditions as exon segments that were predicted for both polyA(+/-) samples (i.e., ones that overlapped between samples), but did not overlap to predictions for any polyA(+) sample (Fig. 3.24 A).

Experimental Validation

See Laubinger et al. [97] for RNA extraction, cDNA synthesis and PCR protocols and primer sequences used.

Computation of Transcribed Fragments (Transfrags)

As an independent method to compare transcriptional activity between polyA(+) and polyA(+/-) samples, we computed transfrags as described by Kampa et al. [85] and implemented in the Affymetrix Tiling Analysis Software version 1.1 build 2. In order to select optimal parameters, we evaluated transfrags generated for root tissues for 900 different combinations of parameters in comparison to annotated genes as detailed above. As optimal setting for all transfrag computations we chose the one with maximal recall at a precision similar to mSTAD predictions (bandwidth 100, signal threshold 6, min run 100, max gap 40; see Fig. 3.19). Among non-repetitive transfrags (at most 25% repetitive probes) comprising at least 4 probes and without overlap to annotated transcripts, the

[7] `http://gbrowse.weigelworld.org/cgi-bin/gbrowse/attax/`

[8] downloaded from TAIR `http://www.arabidopsis.org`, on 15 August 2007

ones specific to polyA(+) or polyA(+/-) samples were computed the same way as for high-confidence mSTAD predictions (Fig. 3.24 B).

2.6.7 Generating TARs for Stress Data

This section describes mSTAD's application to tiling array data from stress treated *A. thaliana* seedlings (see Table 4.4), the identification of TARs showing stress-induced expression patterns and characterizes some properties of these TARs.

Detection of Unannotated Transcriptionally Active Regions (TARs)

Before we detected transcriptionally active regions (TARs) using the mSTAD algorithm, we normalized raw tiling array data applying background correction (see Section 2.3.1; [14]), quantile normalization (see Section 2.3.2; [11]), and finally transcript normalization (see Section 2.3.3). Afterwards we trained mSTAD on 1 h and 12 h mock controls. Genome-wide predictions for 1 h salt, osmotic, ABA, cold and heat stressed samples were made by the models trained on the 1 h mock control; for 12 h of salt, osmotic, ABA, cold and heat stress the models trained on the 12 h mock control sample were used. From the predicted TARs a set of unannotated, high-confidence predictions (referred to as "new TARs") was extracted (as also described in Section 3.3.3), requiring that the TARs included at least 4 probes, fewer than 25% repetitive probes, average expression level 6-10 and an overlap to annotated exons of at most 25 nt.

Testing TARs for Stress-Induced Expression

Each TAR meeting the above criteria of an unannotated high-confidence region was tested for stress-dependent increase in expression level. With the Wilcoxon rank-sum test (also known as Mann-Whitney-U-Test; we used the two-sample version of the Kruskal-Wallis test implemented in the Matlab statistics toolbox) we compared the intensities of all probes inclusive to the TAR of interest between the stress sample and the corresponding mock control (pooling replicate intensities). When the median intensity under stress was significantly higher than that of the control at a p-value of 5%, a TAR was called "stress-induced".

RT-PCR Analysis of New Stress-Induced TARs

RT-PCR validation experiments were performed using the same protocol as in Section 2.6.6.

Overlap Between New TARs Identified under Different Stress Conditions

In a pair-wise comparison of stress-induced new TARs, we counted positions where new TARs induced by different stresses overlapped. Subsequently, we normalized these counts by the total number of nonredundant positions corresponding to new TARs which were induced by either of the two stress conditions to obtain the percentages shown in Fig. 3.26 C.

Assessing Evolutionary Conservation of New TARs

Whole-genome alignments between *Arabidopsis thaliana*, *Oryza sativa*, *Populus trichocarpa* and *Sorghum bicolor* were obtained from the VISTA project[9] [52]. These whole-genome alignments were generated with methods described in Brudno et al. [19], Couronne et al. [33], Kent [88]. As a proxy for conservation of a region of interest, we assessed the number of sequence identities in the alignment corresponding to a TAR. Afterwards, sequence identity counts were normalized by transcript length and the number of aligned species (three). As a control for the new TARs in each stress sample, we randomly sampled 100 times as many annotated exons assessing their degree of conservation in the same manner. Resulting histograms are shown in Fig. 3.27 A for 1 h salt stress and for all other stress samples in Fig. 4.8.

Calculating Distances Between TARs and Neighboring Genes

For each stress-induced new TAR, we determined the distance between its start and the nearest annotated gene upstream as well as the distance between its end and the nearest gene downstream. The histogram shown in Fig. 3.27 B was computed from the minimum of these two distances. A distance of 1 can result either from a small overlap to (an) exon(s) or from the new TAR being located in an intron of an annotated gene (for other samples see Fig. 4.9).

2.7 Incorporating Sequence Information into mSTAD

In this section we describe an extension of the mSTAD method which enables it to exploit features of the genomic sequence, specifically the local sequence context around splice sites, in addition to hybridization measurements. Instead of directly incorporating sequence features we provided mSTAD with pre-computed splice site predictions [170]. A description of the enhanced model is followed by details on evaluating its performance relative to the original method. This is unpublished work done together with Jonas Behr and Gunnar Rätsch.

As a first preprocessing step, genome-wide splice site predictions had to be mapped into the probe-grid defined by the tiling array design. Moreover, we extended mSTAD's state model and re-defined its loss function as detailed in the following. We called the resulting transcript mapping method margin-based segmentation of tiling array data with splice site predictions (mSTADsp).

2.7.1 Splice Sites Predicted from Genomic DNA Sequences

Genome-wide splice site predictions from genomic sequences were computed by Jonas Behr as described in Sonnenburg et al. [170]. Support vector machines with a so-called "Weighted-Degree kernel" [136] classified genomic sequences around known splice sites in order to discriminate them from decoy sites exhibiting the same consensus dinucleotide

[9] `http://pipeline.lbl.gov/downloads.shtml`

("GT" and "GC" for donor and "AG" for acceptor splice sites). Two sets of splice site predictions were generated, one where large sequence windows (as proposed in Sonnenburg et al. [170]) were extracted that contained 81 nt and 60 nt of exonic and intronic sequence, respectively (in the following referred to as "w141"); for the second set of predictions the input sequence was restricted to 40 nt of intronic and 10 nt of exonic sequence (referred to as "w50"), aiming to minimize preferential detection of splice sites in coding sequences.

2.7.2 The State Model Utilized by mSTADsp

The key idea for extending mSTAD's state model is the introduction of a second set of states which allow learning from splice site predictions when segmenting tiling array data. All transitions in mSTAD's original model where replaced by an additional splice site state that was connected by newly inserted transitions to the adjacent hybridization states (Fig. 2.7). On the basis of this interleaving of splice site states and hybridization states, learning from hybridization and sequence features was possible within mSTAD's existing training and prediction framework. To compensate for the increase in model complexity upon the introduction of splice site states, only a single exon state and a single intron state were used per expression quantile, whereas mSTAD's model contained a pair of each of these (Figs. 2.5 and 3.20).

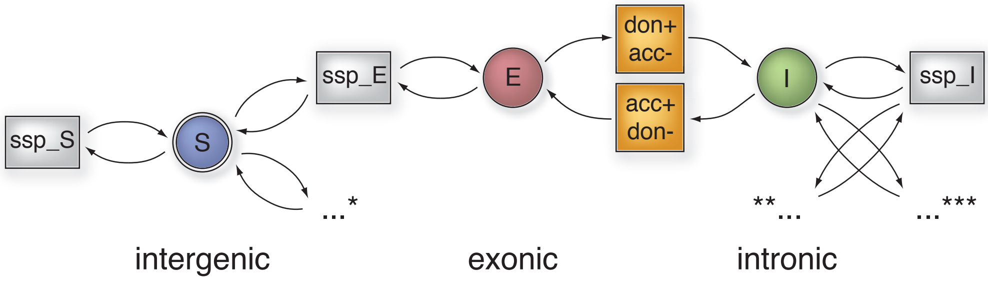


Figure 2.7: Sketch of the state model employed in mSTADsp. Shown is a submodel with exon and intron states for the first of $Q = 10$ discrete expression levels. States corresponding to hybridization signals are indicated by circles (S, E, I) and splice site states by rectangles (don, acc, ssp_S, ssp_E, ssp_I). For transcripts originating from the Watson strand, a strong splice donor signal (don+) is expected between exon and intron, similarly a strong acceptor signal (acc+) between intron and exon probes. When segmenting a gene on the Crick strand with this strand-insensitive model, transitions between exon and intron correspond to splice acceptors (acc-), whereas intron-exon boundaries feature splice donors (don-). Ideally, there are no strong splice signals at the remaining splice site states (ssp_S, ssp_E, ssp_I). Transitions into submodels for the next expression level are indicated by "..." and lead to the following states (not shown here) * — $\mathsf{ssp\_E}_2$, ** — I_2, *** — $\mathsf{ssp\_I}_2$.

2.7.3 Deriving Features for mSTADsp

As a first feature for mSTADsp we used exactly the same hybridization signals as in mSTAD. Additional features captured information on the start and end of internal exons.

Originally, splice site predictions were available for each consensus dinucleotide on Watson and Crick strand (for the formal definitions below we assume a prediction score of $-\infty$ for all other sites). The exon start feature x_{ie} between probes i and $i+1$ with center locations p_i and p_{i+1}, respectively, was derived as follows:

$$x_{ie}(i) = \max(\{a^{(+)}(q) \mid p_i \leq q < p_{i+1}\} \cup \{d^{(-)}(q) \mid p_i \leq q < p_{i+1}\}).$$

where $d^{(+)}(q)$ and $d^{(-)}(q)$ denote a splice donor prediction score at position q on Watson and Crick strand, respectively. By $a^{(+)}(q)$ and $a^{(-)}(q)$ we denote the respective splice acceptor prediction scores. The exon start feature was evaluated in the acc+ / don- state (Fig. 2.7). In the don- / acc+ state, an exon end feature x_{ei} was evaluated which was derived analogously:

$$x_{ei}(i) = \max(\{d^{(+)}(q) \mid p_i \leq q < p_{i+1}\} \cup \{a^{(-)}(q) \mid p_i \leq q < p_{i+1}\}).$$

Finally, we employed a fourth feature, $x_{ns}(i) = \max(\{x_{ie}(i),\ x_{ei}(i)\})$. This feature was evaluated in states denoted ssp_S, ssp_E or ssp_I (Fig. 2.7). Because within an exon, intron, or intergenic region, ideally no strong splice signal should be encountered at all, we added this feature to facilitate learning a feature scoring function that penalized the occurrence of *any* of these cases.

2.7.4 Regularization and Loss

For mSTADsp we used the same quadratic regularizer as described before (see Section 2.6.4), but additionally constrained the feature scoring functions of the splice site states to be monotonic, considering that non-monotonic transformations of the SVM prediction scores would be exclusively due to the stochastic nature of the learning process. This was achieved by adding constraints of the following form to the optimization problem introduced in Section 2.4.4:

$$\begin{array}{lcll} |\theta_{j,k,l} - \theta_{j,k,l+1}| & \leq & 0 \quad \forall\, l = 1, \ldots, S-1 & \text{for monotonically increasing functions } g_{jk} \\ |\theta_{j,k,l} - \theta_{j,k,l+1}| & \geq & 0 \quad \forall\, l = 1, \ldots, S-1 & \text{for monotonically decreasing functions } g_{jk} \end{array}$$

for some feature j and state k; S denotes the number of supporting points of the piecewise linear feature scoring function g_{jk}. For x_{ns} monotonically decreasing feature scoring functions were learned, whereas scoring functions for x_{ie} and x_{ei} were constrained to be monotonically increasing. Moreover, constraints enforcing similar scoring functions for the same splice site feature in corresponding states of different expression levels were added (as described in Section 2.6.4).

The loss function employed in mSTAD was extended by additional penalties for confused splice site states. All splice site state confusions incurred a position-wise loss of 1 that was added to the path-loss Δ (see Section 2.6.4).

2.7.5 Performance Assessment

Augmenting the example sequences of hybridization signals used for mSTAD with additional splice site features as well as partitioning examples into training, validation and test set in exactly the same manner helped to minimize biases when comparing the performance of mSTAD and mSTADsp.

For evaluation purposes, a set of 1,000 full-length cDNA-confirmed genes annotated in TAIR7 [178] was used and prediction accuracy was assessed for test predictions generated in a two-fold cross-validation procedure. Two disjoint training sets, each containing 1000 sequences, were sampled around annotated genes such that they were also disjoint from the respective test set. Two instances of mSTAD as well as two instances of mSTADsp were trained on these sets. Both, mSTAD and mSTADsp, were retrained using maximum likelihood estimation to obtain mSTAD HMM and mSTADsp HMM (see Section 2.4.7). For mSTADsp, training and evaluation was repeated once more on corresponding data based on the second set of splice site features to obtain mSTADsp w141 and mSTADsp w50. Precision and recall were calculated with the same routines as for the evaluation of mSTAD (see Section 2.6.5 and Fig. 3.19). Finally, precision and recall estimates were corrected for biased expression of test genes in order to more closely reflect the accuracy expected for genome-wide predictions. Instead of directly assessing precision and recall across all test examples, these measures were calculated separately for each discrete expression level. Afterwards we averaged them across levels and thereby effectively down-weighted example sequences around genes with medium to high expression level that were overrepresented in the test set (Fig. 3.32).

2.8 Chapter Summary

Whole-genome tiling arrays hold great promise for many biological applications, yet the analysis of the resulting hybridization data poses many challenges. We developed a normalization pipeline to correct for several biases and sources of signal variability (Sections 2.3.1, 2.3.2, 2.3.3). Its most important step, a novel transcript normalization method, addresses the well-studied problem that the sequences of tiling probes themselves have a strong influence on their hybridization properties and thus on the resulting signal. As a consequence, signal variation due to divergent probe sequences impedes comparisons between different probes and therefore also subsequent analyses, such as transcript identification. We took a novel modeling approach to directly reduce the variance of observed signals from an ideally expected, constant signal for all probes with the same concentration of bound target molecules. Its benefits will be shown in the following chapter (Section 3.2.5).

We formalized the biological tasks of polymorphic region prediction as well as transcript identification as segmentation (or label sequence learning) problems by extracting meaningful label information and by deriving features useful for learning. To solve these problems, we implemented and engineered Hidden Markov Support Vector Machines (HM-SVMs) [3, 180, 184]. This included the design of an expressive linear feature map, problem-specific state models, loss functions and regularizers (Sections 2.4.3, 2.5.3, 2.5.6, 2.6.2, 2.6.4). For the problem of *de novo* transcript identification from tiling arrays, we conducted a comprehensive empirical assessment of the properties of discriminative HM-SVMs in comparison to generative HMMs. We observed that HM-SVMs, albeit costly to train in terms of CPU time, made significantly more accurate test predictions and were much more robust to label noise, highly desirable properties for the analysis of biological data. Importantly, we found that the performance difference between HM-SVMs and HMMs could largely be attributed to the careful design of the HM-SVM loss function — for a simple, problem-independent loss, HM-SVMs and HMMs showed almost identical accuracy (Section 2.4.8).

For the first time, we systematically identified polymorphic regions including deletions from resequencing microarrays taking a rigorous approach based on HM-SVMs (Section 2.5). Since no other computational methods were available for comparison, we carefully assessed our predictions against other data sources to verify their accuracy (results will be shown in Section 3.1).

For the better-studied problem of transcript identification from tiling arrays, we developed a new HM-SVM-based method, which introduced several novelties in the modeling approach (Section 2.6). Most importantly, we maintain different parameter sets for genes with different expression levels in the form of an elegant state model. Moreover, we show that this model can easily be extended to exploit genomic sequence features in addition to hybridization features (Section 2.7). The existence of other methods for transcript identification [72, 77, 85, 120] allowed us to compare their performance to ours (results will be shown in Section 3.3.2).

3 Results and Discussion

The first part of this chapter describes our analyses of resequencing microarrays in order to identify highly polymorphic regions in the *A. thaliana* genome. We then turn to the analysis of transcriptome tiling arrays and present the results of our transcript normalization method. Subsequently, we describe the results of our approach to transcript identification including (i) an evaluation of its performance in comparison to other methods, (ii) the biological discoveries resulting from its application to tiling arrays profiling the transcriptome of *A. thaliana* tissues and developmental stages and (iii) its application to a tiling array-based survey of transcriptional stress response in *A. thaliana*. The last results section covers the extension of our hybridization-based transcript identification method to one which also exploits genome sequence information. The chapter is concluded by a discussion about extending the developed transcript normalization and identification methods such that they are applicable to RNA-seq, a sequencing-based assay of transcriptional activity, which is currently revolutionizing transcriptomics.

3.1 Margin-Based Prediction of Polymorphic Regions (mPPR)

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 [29]. However, where multiple SNPs or insertion/deletion (indel) polymorphisms are closely adjacent (occur within the same 25-mer), all oligonucleotide probes interrogating this local context harbor off-center mismatches, and SNP prediction is generally not possible. For such regions, hybridization is suppressed for contiguous features in a tiling path. This pattern is therefore a signature of high underlying polymorphism, either in the form of closely linked SNPs or small indels, or potentially of larger deletions (Fig. 3.1 A,B,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 may be (i) monomorphic to the reference sequence, or alternatively may be (ii) so dissimilar that no underlying polymorphisms are detected.

In this section, which is based on joint work with Richard M. Clark, Korbinian Schneeberger, Anja Bohlen, Detlef Weigel and Gunnar Rätsch (see p. 137 for author contributions) [203], we describe a novel machine learning method, suitable for detecting tracts of high polymorphism from resequencing array data. Formally, the prediction task is to label each tiled position in the genome as either (i) conserved or (ii) at or immediately adjacent to a polymorphism (Fig. 3.1 D). Here, we define *polymorphic regions* (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 (Fig. 4.2 for a discussion of these distances).

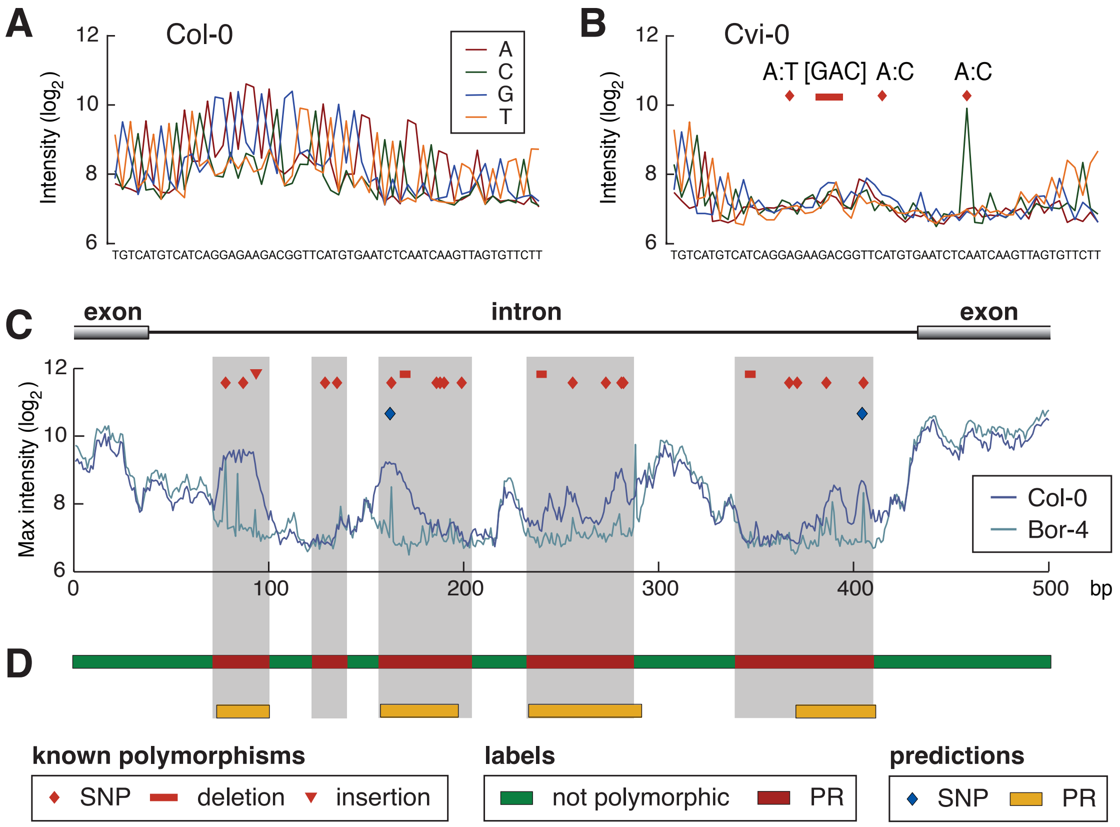


Figure 3.1: Effect of polymorphisms on hybridization patterns, labels for the mPPR algorithm, and polymorphic predictions. **(A)** Log_2 intensities for oligonucleotides in a 56 bp tiling path (chromosome 4, positions 8,375,747 to 8,375,802) for the reference Col-0 accession. Intensities for each sequence (see inset) are given and are averages for the forward and reverse strand features tiled on the arrays (see Section 2.5.5). **(B)** Corresponding data from accession Cvi-0 for which 3 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 3 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 to 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 2 of the 21 known Bor-4 polymorphisms (17 of which are SNPs) were predicted in MBML2. **(D)** The corresponding 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. Plotted data are from Clark et al. [29], Nordborg et al. [128].

Our method, which we call **m**argin-based **P**rediction of **P**olymorphic **R**egions (mPPR), employs HM-SVMs to accurately identify PRs from resequencing array data of the reference plant, *Arabidopsis thaliana*, where SNP polymorphism is higher than for human [192, and references therein], and for which indel polymorphisms are common [73, 128].

We applied mPPR to a previously published *Arabidopsis* resequencing array data set, hereafter called “AtAD20” [29]. It contains data generated for $> 99.99\%$ of bases in the

119 Mb reference genome [73] for each accession [29]. The data were previously used to identify ≈ 648,000 SNPs at a precision of about 98% (the "MBML2" SNP data set). SNPs from highly polymorphic regions were underrepresented among these SNP predictions, highlighting that such regions present a substantial challenge. Likewise, a fine-scale, genome-wide map of insertions and deletions (indels) in *Arabidopsis* was lacking, as precise methods for the identification of indels had not been developed. However, clustered polymorphisms and indels, which can comprise more than 15% of polymorphisms in eukaryotic genomes [e.g., 36, 116, 190], are a central component of sequence variation, and contribute to phenotypic variation. With mPPR, on average ≈ 288,000 polymorphic regions were predicted per accession at a precision of about 97% revealing a large proportion of polymorphisms absent from the MBML2 data set. 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.

3.1.1 Known Polymorphisms for Training and Evaluation

The mPPR 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, 1,213 fragments of ≈ 550 bp in length and located throughout the genome had been sampled by PCR and dideoxy sequencing [128]. This data set, hereafter called "2010", covers about 0.5% of the genome per accession, and harbors ≈ 2700 SNPs and ≈ 400 indel polymorphisms per target accession [128]. Col-0, the reference accession, was included in the AtAD20 accession set [29], 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 [73]). Our method also used information about the repetitiveness of each arrayed 25mer oligonucleotide determined from the Col-0 reference sequence (see Section 2.1 and [29]). In particular, we separately modeled repetitive sequences from non-repetitive sequences in an effort to avoid fragmentation of predictions in regions of low to moderate repeat content (see Sections 2.5.3 and 2.5.5).

We trained our method on 60% of the 2010 data, used 20% for hyper-parameter tuning, and 20% for evaluation; we employed a 5-fold 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 as a false negative (FN). We used these counts to assess precision and recall[1] defined as $\mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$ and $\mathrm{TD}/(\mathrm{TD}+\mathrm{FN})$,

[1] which is the same as sensitivity

respectively (Fig. 3.2 for details). 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 precision and recall (Fig. 3.3 A, and Section 2.5.7 for details). For 2010, we generated predictions at a precision of $\geq 90\%$ for $\lambda = 75\%$ (Fig. 4.3 for the effect of varying λ on precision and recall). Across all sequence types and accessions, our method identified 56% of PRs in 2010, and performance estimates varied only moderately between accessions (Table 4.2). In *Arabidopsis*, coding sequences have higher GC content and sequence complexity than noncoding sequences [73]. These factors are favorable for hybridization-based methods [29, 99], and likely contributed to the higher recall rate in coding regions (e.g., about a 1.3-fold difference compared to non-coding sequences at a similar precision; see Fig. 3.3 A and Table 3.1). As minor differences in prediction boundaries affect performance estimates — especially for small predictions — we also assessed the performance of the predictions with a relaxed overlap criterion (Fig. 4.3). For $\lambda = 50\%$, recall was slightly higher, and precision was at least 95% for all sequence types and $\approx 97\%$ on average (Table 3.1).

	Coding	UTRs + introns	Intergenic	2010	Genome
$\lambda = 75\%$					
Precision	92.6%	88.8%	88.3%	90.4%	89.3%
Recall	63.6%	50.9%	49.1%	55.6%	52.0%
$\lambda = 50\%$					
Precision	97.4%	97.9%	95.8%	97.2%	96.6%
Recall	65.5%	54.4%	51.7%	58.2%	54.7%

Table 3.1: Precision and recall for PR predictions assessed with 2010 for different overlap cutoffs, λ (see main text). The relative abundance of sequence types differs between 2010 and the whole genome [29], and precision and recall were re-estimated accordingly for the whole genome predictions (column "Genome", see Section 2.5.7).

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 (Table 3.2). 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, about 90% of bases within predictions were at or within 6 bp to a known polymorphism (Fig. 3.3 B).

While PR predictions typically reflected the underlying patterns of polymorphisms with high accuracy, prediction boundaries sometimes differed substantially from labels, and for

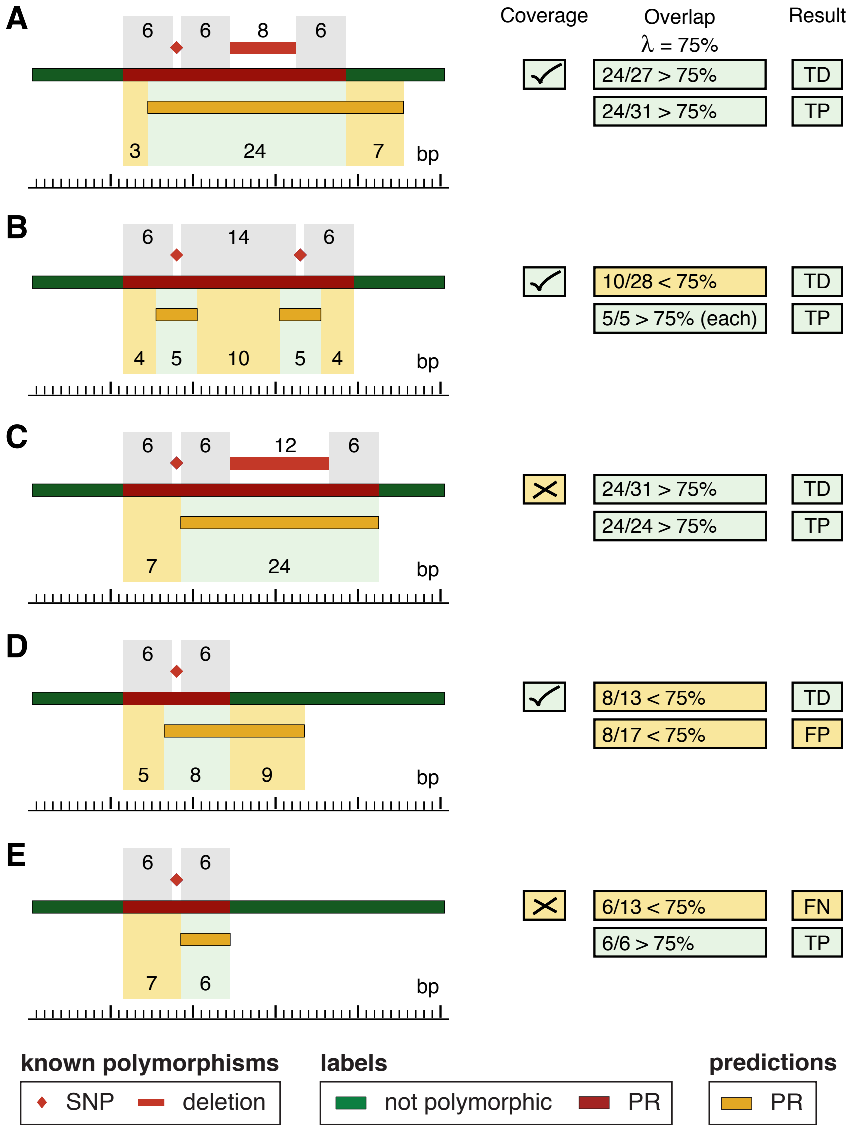


Figure 3.2: Illustration of performance assessment. To calculate recall, i.e., for whether each label PR was a true discovery (TD, green shading) or a false negative (FN, yellow shading), we first checked if all underlying polymorphisms were included in one or more PR predictions (boxes with title "coverage"). 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**). Precision 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, yellow shading, as in **D**).

some regions even highly clustered polymorphisms were not identified (Fig. 3.1 C,D). In large part, such false negatives occurred for regions with poor hybridization properties in the reference accession (e.g., compare predictions to reference feature intensities for regions (2) and (5) in Fig. 3.1 C,D; see also Fig. 4.4). Additionally, although explicitly modeled by our method, repeats were overrepresented among false negative predictions. For example,

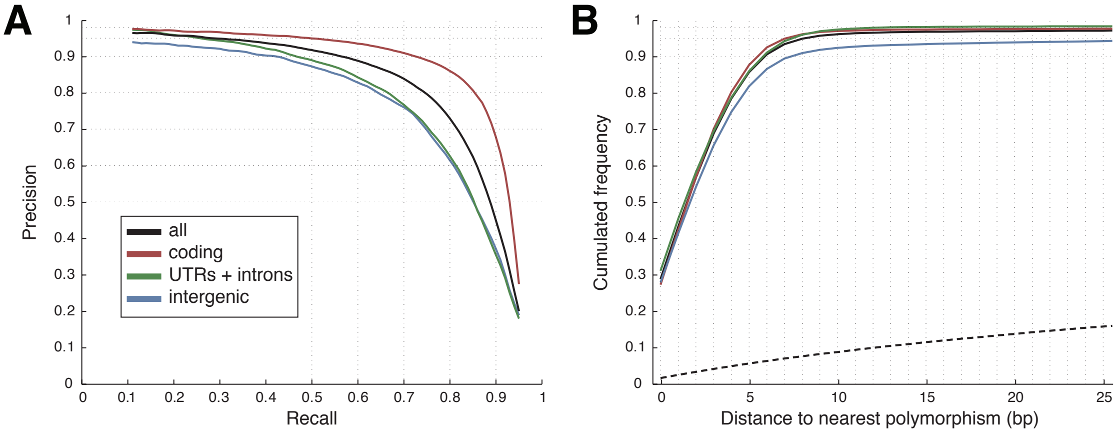


Figure 3.3: Relationship between precision and recall for PR predictions with overlap criteria $\lambda = 75\%$. **(A)** Precision-recall curves averaged over cross-validation test subsets for different sequence types (see inset for color code). PRs that contained more than one sequence type were assigned to the type comprising the majority of the prediction. **(B)** Precision 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 6 nucleotides to a known polymorphism. The dashed black line indicates the relationship expected by chance (that is, predictions were assigned to random genomic locations for calculating distances).

	Bases (kb)	PRPs	Single-SNP	Multi-SNP	Deletion	Insertion	Complex	Empty
2010	10,967	20,073	12,435	4,584	438	242	1,607	767
C24	14	65	27	23	4	0	9	2
Cvi-0	37	169	76	60	6	2	23	2
L*er*-1	37,871	76,020	41,052	18,331	2,257	1,319	9,771	3,290

Table 3.2: 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").

in 2010 5.5% of all positions were repetitive (see Section 2.1), while the fraction of repetitive positions in false negative 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.

3.1.2 Prediction Content and Comparison to SNP Calls

We designed mPPR to produce predictions that complement existing SNP data sets ascertained from resequencing array data (Fig. 3.4). Although our method only identifies the approximate location of polymorphisms, 74.8% of clustered SNPs ($\leq$ 18 bp away from the nearest polymorphism) in 2010 were included within boundaries of PR predictions

(Table 3.3). 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 to 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 7- to 10-fold as many SNPs as MBML2 (Table 3.3).

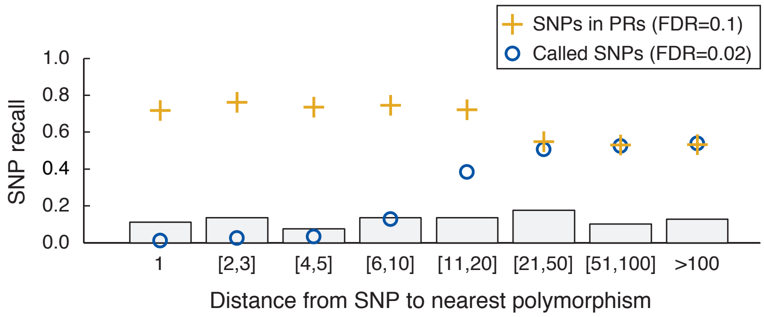


Figure 3.4: Dependency of SNP recall on distance between polymorphisms by detection method. SNPs were partitioned according to the distance to the nearest polymorphism. The frequency of SNPs in each distance bin (x-axis) is shown as bars. Recall rates per distance category are given for MBML2 SNP calls (circles) and inclusion within PR prediction boundaries (crosses).

	Coding		UTR + intron		Intergenic		All	
	PRs	MBML2	PRs	MBML2	PRs	MBML2	PRs	MBML2
Clustered SNPs	85 (65)	21	71 (64)	7	66 (57)	9	75 (63)	12
	[9196]		[10793]		[5608]		[25597]	
Isolated SNPs	61 (14)	69	54 (22)	41	48 (22)	37	55 (18)	53
	[10294]		[6774]		[5870]		[22938]	
All SNPs	72 (38)	46	64 (48)	20	57 (39)	23	66 (42)	31
	[19490]		[17567]		[11478]		[48535]	

Table 3.3: Recall by polymorphism and sequence type. For MBML2 (precision $\approx 98\%$ [29]), 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 (precision $\approx 90\%$; see Table 3.1), 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.

3.1.3 Whole-Genome Predictions and Evaluation

HM-SVMs trained on 2010 data were used for genome-wide prediction on AtAD20 accessions using the same settings as for evaluations on 2010 data (Table 3.1). Non-redundantly, 27% of the *Arabidopsis 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 (Table 4.2). The accession with the most predictions, Cvi-0, was known from earlier work to be highly dissimilar to Col-0 [128, 150]. By sequence type, intergenic positions were most strongly overrepresented within prediction boundaries (Fig. 3.5).

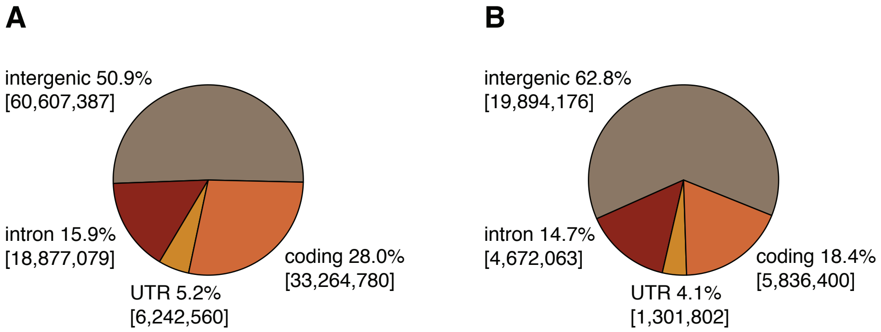


Figure 3.5: **(A)** Arrayed bases by sequence type. **(B)** Non-redundant bases included in PRs by sequence type.

Given the size and genome-wide sampling for the 2010 data [128], our performance evaluations likely generalize well for much of the genome. Nevertheless, the 2010 data is 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 3.1). However, non-coding 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, precision was 96% and 100% (for $\lambda = 50\%$) at a recall rate of 67% and 45% for Cvi-0 and C24, respectively (Table 3.4). Moreover, we assessed our predictions using the much larger 2-fold draft shotgun sequence data available for L*er*-1 (see Section 2.5.8). 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 Section 2.5.8), the prediction quality assessed with 37.9 Mb of aligned sequence data was found to be very similar to that assessed with the 2010 test data: Among the L*er*-1, Cvi-0, and C24 data sets, precision, which is not expected to be strongly affected by errors in the genomic sequence data, varied comparatively little (Table 3.4 and Table 3.5). However, at the first glance recall 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 recall rates, 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 (Table 3.5). The large disagreement between the two sets of PR labels, as well as the discrepancy between recall 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 L*er*-1 data. We therefore multiplied the recall estimate for L*er*-1 predictions obtained from the genomic data by the resulting fold difference in recall estimates for predictions evaluated on 2010 and on the genomic sequences for which the data sets overlapped (a factor of about 1.5; Table 3.5). Both the uncorrected (u) and corrected (c) estimates for recall for the genome-wide L*er*-1 predictions are given in Table 3.4. We thus concluded that 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 than observed for 2010, Table 3.2).

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

Table 3.4: 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. Precision ("Prec.") and recall are given for two different overlap cut-offs, λ (see Section 2.5.8). (a) Precision and recall values were directly compared to the alignments of the L*er*-1 contigs to the Col-0 reference sequence. (b) We corrected the recall rate by a factor estimated from the discrepancies of the recall rates in regions where the 2010 L*er*-1 sequences overlap to the L*er*-1 contigs (see Section 2.5.8 and Table 3.5).

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 uncertainties in the draft genomic data (see Section 2.5.8). Long deletions pose a challenge for our learning method as they were absent from the 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 (Fig. 3.6 B for an example). More than 100 known deletions of greater than 300 bp had been previously characterized

	$\lambda = 75\%$		$\lambda = 50\%$	
	Precision	Recall	Precision	Recall
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 3.5: 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.

in AtAD20 accessions [29], or were characterized in the current study (see Section 2.5.9). These deletions were almost entirely included within PR predictions (Table 4.3, Fig. 3.6 B, and Fig. 3.7).

Finally, we note that extended tracts of repetitive sequences (> 500 bp) are entirely absent from our evaluations. Nonetheless, such sequences are common in *Arabidopsis*, 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. [182] 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. 3.6 A; [29, 182]). 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 *Arabidopsis thaliana* euchromatic genome.

3.1.4 Polymorphism Patterns Ascertained with PR and SNP 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 (Figs. 3.8 and 4.5). Polymorphism patterns apparent in the PR data also resembled that for pair-wise nucleotide diversity as previously calculated with MBML2 [29], as well as for several data sets generated by dideoxy sequencing [29, 128, 151]. Moreover, the patterns were also similar to those observed in single feature polymorphism data collected

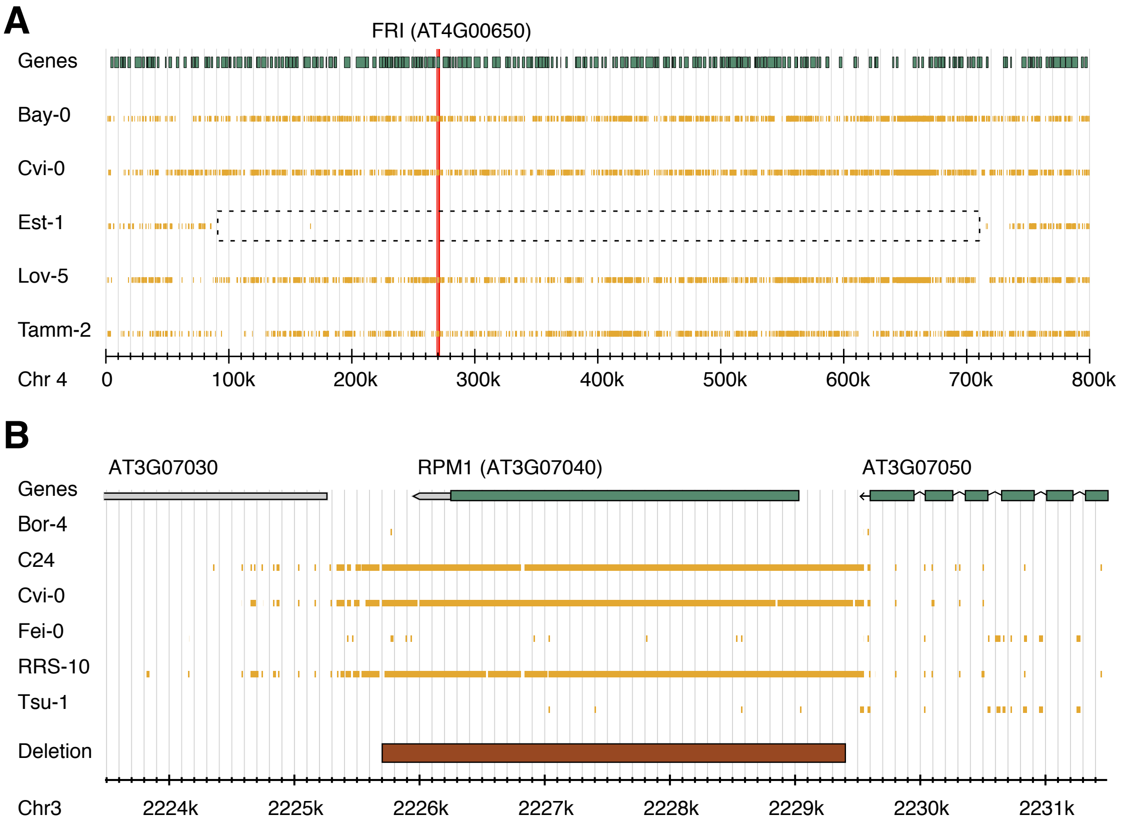


Figure 3.6: PRs reveal haplotype sharing at chromosomal and local scales. **(A)** Genes (green boxes at top) and PRs (yellow blocks beneath) for 5 accessions for 0.8 Mb surrounding the *FRI* locus. In Est-1 a region of about 0.6 Mb (dashed black box) including *FRI* (vertical red line) has been reported to be nearly identical to the Col-0 reference sequence, but divergent in the other accessions shown [29, 128]. Only few 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 *Arabidopsis thaliana* population is as indicated (brown box at bottom) [60, 160]. 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 Fig. 3.7).

with the *Arabidopsis* ATH1 microarray [15]. 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. 3.8).

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 3.3; [29]). 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. 3.9 A). For the PR data, the observed pattern is consistent with the general

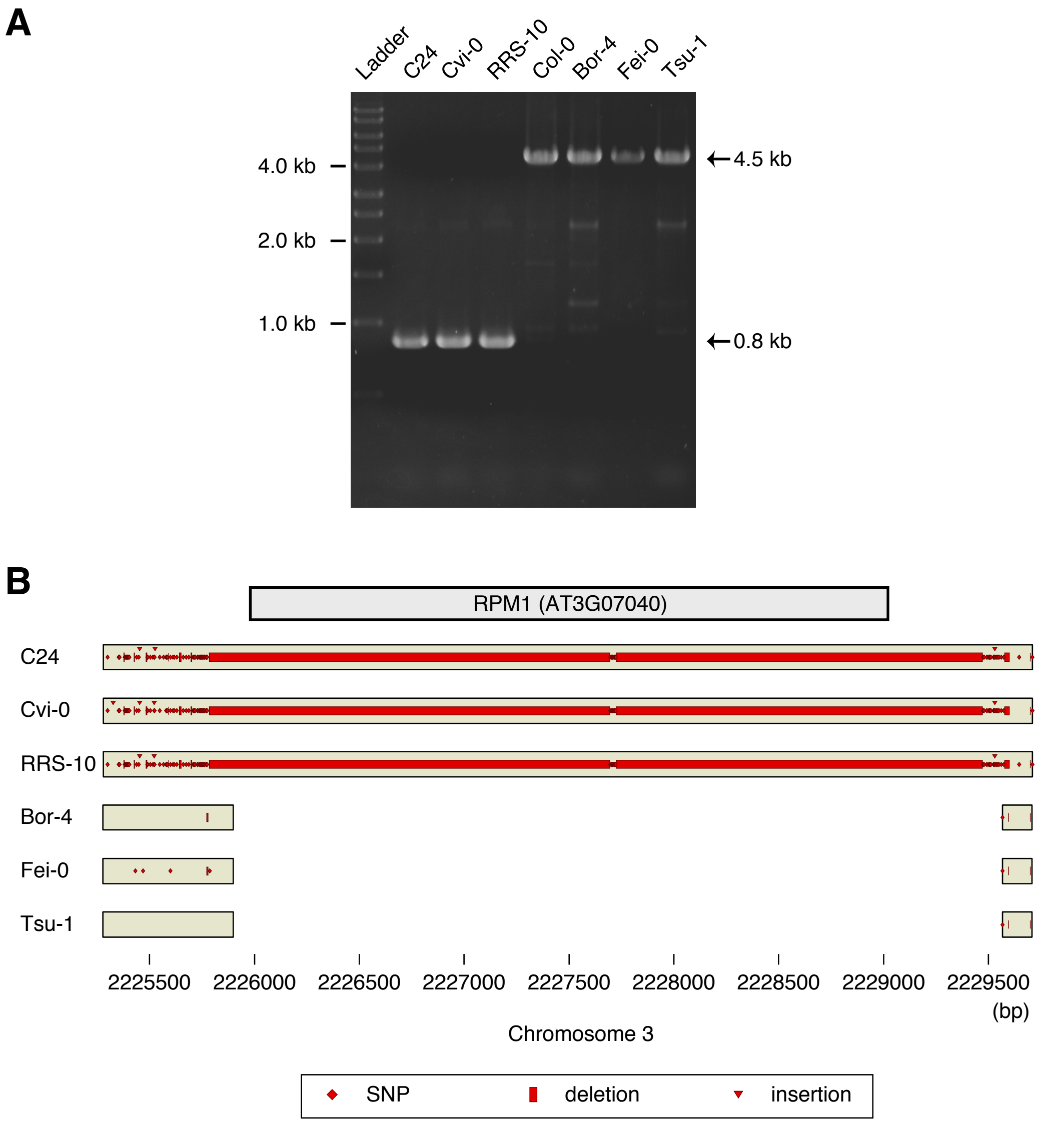


Figure 3.7: Underlying polymorphism at locations of PR predictions at the *RPM1* locus. **(A)** Image of an agarose gel showing PCR products for seven accessions using primers flanking *RPM1* (Lane 1: DNA ladder; Lanes 2-8: PCR products for accessions as indicated at top). Products for three accessions (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. 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 (Fig. 3.6) 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 also Fig. 3.6).

expectation of reduced evolutionary constraint for non-translated sequences, as well as with estimates of nucleotide diversity from 2010 [128]. In addition, inclusion of indels

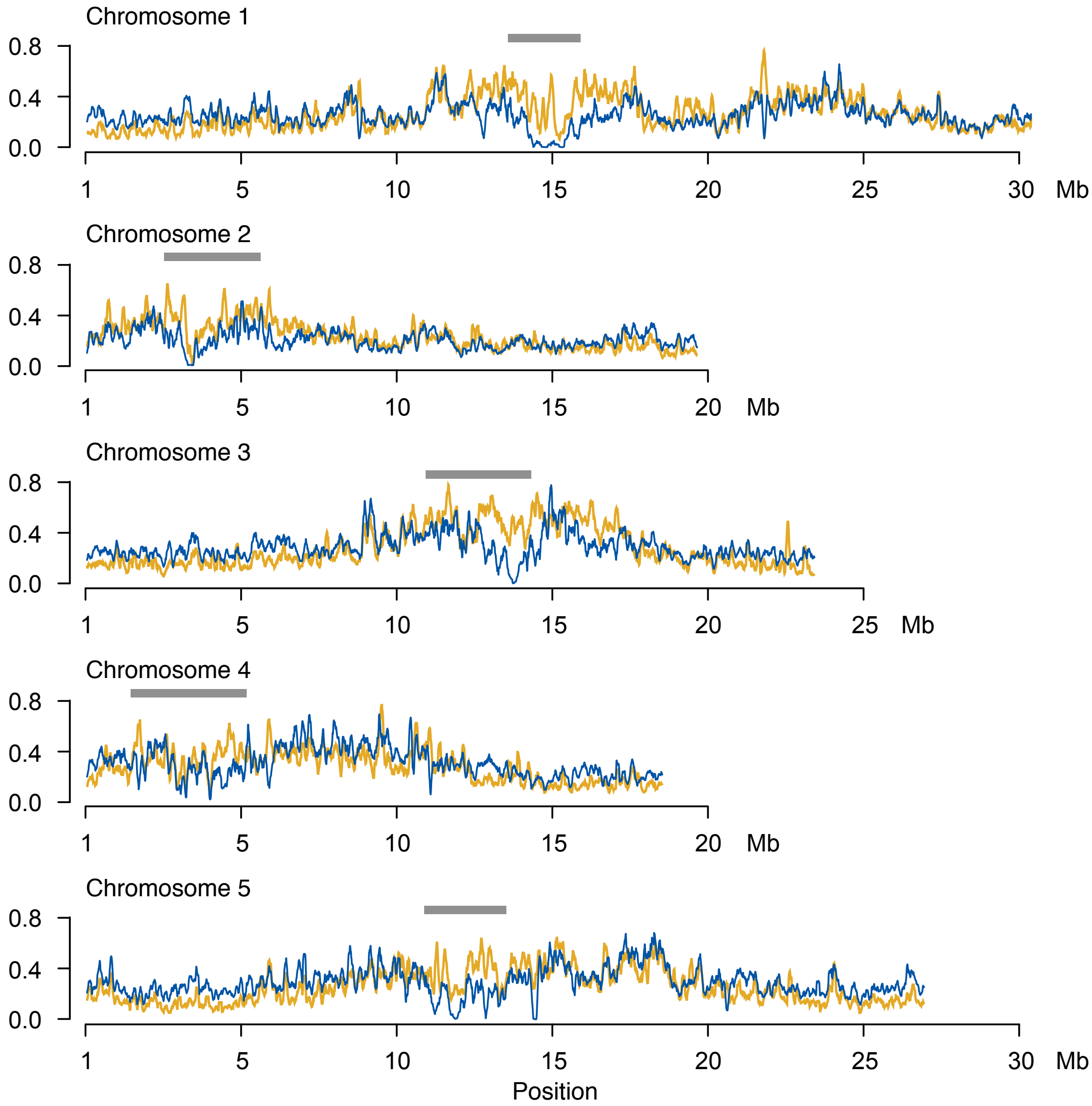


Figure 3.8: Genome-wide patterns of polymorphism in PRs and MBML2 SNPs. A sliding window of 100 kb was used, with values for every 10,000th position plotted. The y-axis displays the fraction of bp in each window included within PRs non-redundantly over all accessions (yellow line). To facilitate visualization, the analogous measure for the SNP data was multiplied by 50 (blue line), and the two measures of polymorphism are broadly correlated (Fig. 4.5). Thick grey bars indicate the approximate positions of centromeres as defined by repeat content (see Section 2.1, and [29]).

as prediction targets for mPPR, coupled with the bias for indel polymorphisms in non-coding regions [73], 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 recall for MBML2 is very low (Table 3.3; [29]), and for which diversity estimates from 2010 are largely limited to sequences near genes [128]. Average levels of polymorphism varied as a function of distance from coding sequences, and were asymmetric relative to gene orientation (Fig. 3.9 B). Extending upstream to 5' UTRs, polymorphism reached a plateau at about 450 bp, while the analogous plateau was reached within about 50 bp downstream from 3' UTRs. Upstream to transcription start sites, polymorphism tended to be inversely associated with the density of predicted *cis*-regulatory elements [129]. 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. 3.9 C; Zeller et al. [supplement of 203], [129]). This observation is unlikely to result from an artifact in the PR data; a similar pattern is apparent in an inter-specific comparison of promoter regions between *Arabidopsis thaliana* and a close relative, *Boechera stricta* [191]. 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 under-representation of overlaps between PRs and transcriptional *cis*-elements predicted in a previous study [129]. This finding indicates that in *Arabidopsis* 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., 100]. An implication of this observation is that deep sampling of variation within *Arabidopsis thaliana* populations will be important for both detecting *cis*-regulatory sequences and for characterizing their evolution.

3.1.5 Highly Polymorphic Genes and Gene Families in Arabidopsis

At the local scale, we used PR predictions to characterize, at high resolution, genes that are highly polymorphic in the *Arabidopsis 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, intra-specific patterns of genic polymorphism predicted inter-specific conservation, with lower PR content for *Arabidopsis* genes with orthologs in black cottonwood *(Populus trichocarpa)*, the most closely related plant with a sequenced genome [185] [supplement of 203]. Among large gene families within *Arabidopsis thaliana* ($n > 125$; [29]), variation in PR content was readily apparent (Figs. 3.10, 3.11 and 4.6). Generally, gene families with many members affected by SNPs expected to impact on gene function ("large-effect SNPs", [29]) also tended to have relatively high levels of PRs in coding regions (Fig. 3.11). Transcription factors, for which MBML2 SNP data suggested strong purifying selection, harbored few members with high PR content (Figs. 3.11 and 4.6). In contrast, higher PR content was observed for F-box genes (Figs. 3.11 and 4.6), for which many inactivating mutations have been identified [29], and for which patterns of sequence variation indicate high death rates in the *Arabidopsis thaliana* genome [181]. Among large gene families, nucleotide-binding leucine rich repeat (NB-LRR) genes that mediate disease resistance exhibited extreme levels of polymorphism (Figs. 3.10 and 3.11), a finding that was even apparent in low resolution predictions of polymorphic regions from AtAD20 data [29].

Characterizing polymorphisms in transcribed *Arabidopsis* sequences at high resolution,

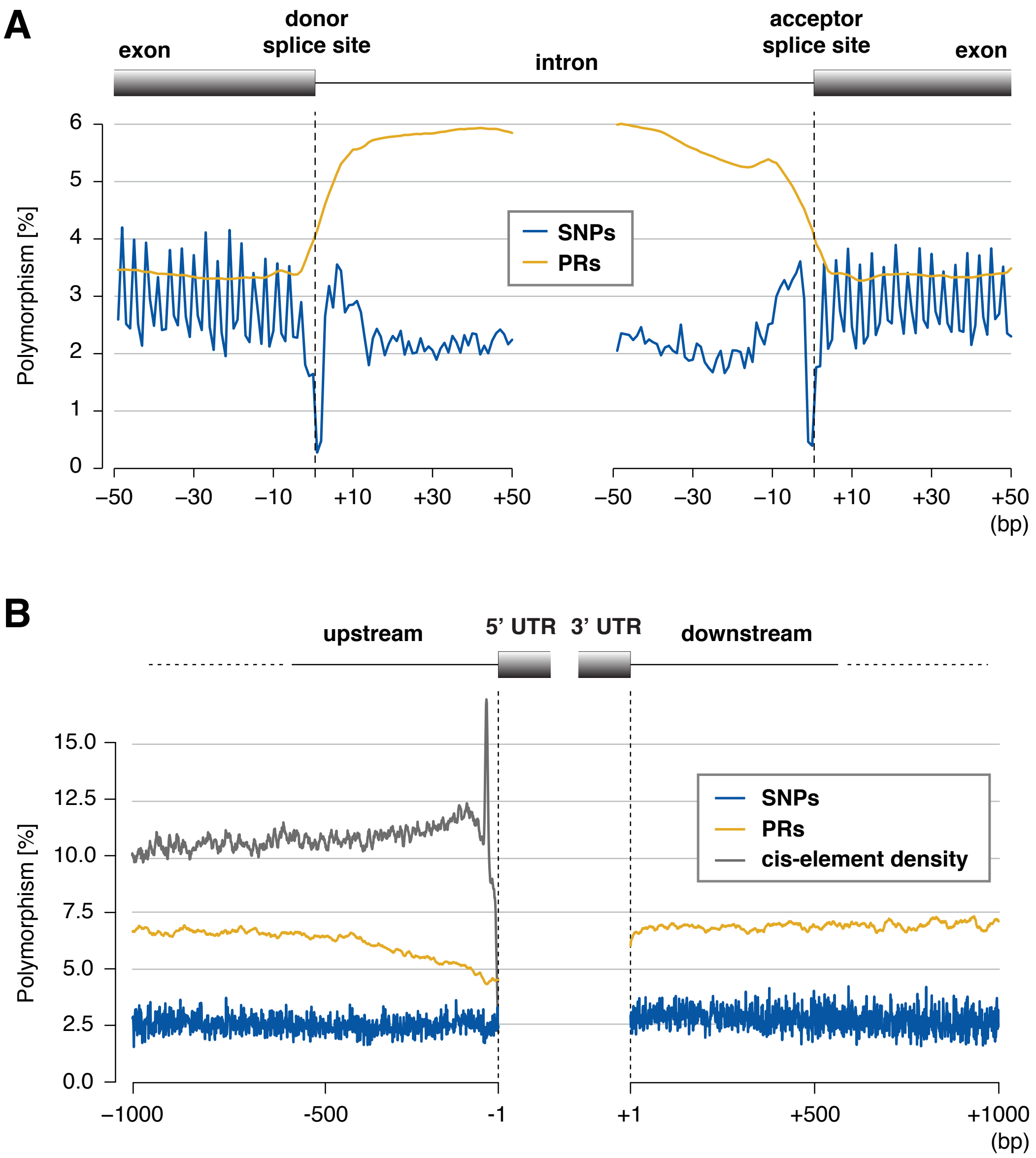


Figure 3.9: Patterns of polymorphism apparent in PR and SNP data in noncoding regions. **(A)** Polymorphism near splice donor (left) and splice acceptor (right) sites as averaged over 116,971 splice sites and assessed with both the PR prediction and MBML2 (SNP) data sets (see inset, and Section 2.5.12 for details of polymorphism estimation). 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 1,000 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.

we found hundreds of transcribed regions containing or even largely covered by PRs in one or more accession. As these represent genes from many families, the evolutionary

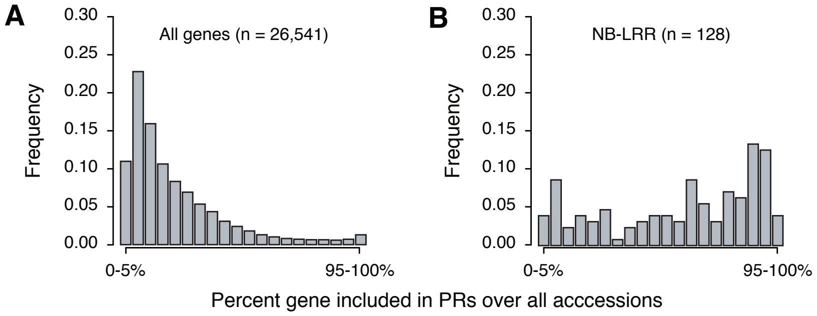


Figure 3.10: Percent of coding NB-LRR genes included in PRs over all accessions in comparison to all genes. Distribution of coding genes as a function of percent inclusion in PRs for all genes **(A)** and NB-LRR genes **(B)**, respectively (see Section 2.5.14).

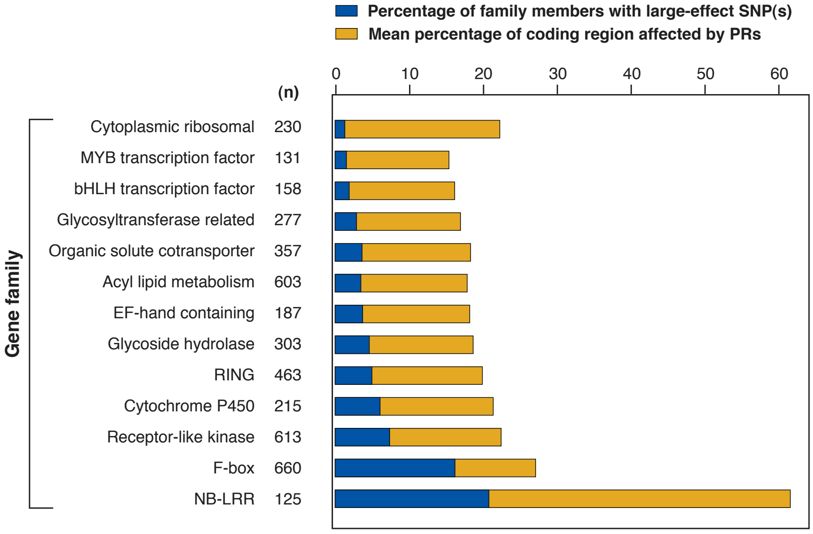


Figure 3.11: SNPs expected to impact on gene structure ("large-effect SNPs" [29]) in comparison to PRs affecting coding sequences by gene family (for families with $n \geq 125$ members). Bar length corresponds to the percentage of family members affected by one or more large-effect SNPs and to the percentage of coding region inclusive to PRs averaged over family members, respectively.

forces shaping patterns of genic polymorphisms are expected to be manifold. In some cases, PRs affecting a large proportion of gene loci may reflect the absence of selection at annotated genes that are in fact pseudo-genes. 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 [4, 29], the predominant class of disease resistance (R) genes in plants [81]. In our study, family-wide polymorphism for NB-LRR genes was extreme, as also noted from earlier work with the AtAD20 data [29], as well as from studies of a selected set of NB-LRR genes in *Arabidopsis* [4, 61, 160]. Nevertheless, polymorphism levels for individual NB-LRR genes varied greatly; some genes were almost entirely included in PRs (Fig. 3.6 B), 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 showing little or no variation may have been targets of recent positive selection (sweeps) in *Arabidopsis 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., 166]. The extent to which other highly polymorphic genes identified in this study mediate interactions with the biotic (or potentially abiotic) environment requires empirical study.

As our PR predictions have high precision and recall in non-coding 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 *Arabidopsis* miRNAs with homologs in other species [82], very little variation was observed for the 21 nt miRNA sequences required for miRNA-mediated gene suppression (Fig. 3.12). Marginally higher variation was observed for the complementary miRNA* sequence, while PR content was 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 [48, 135], 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. 3.12). Whether this pattern reflects poor annotation for the non-conserved miRNAs, or potentially the evolution of new genes that are not fixed in the population, remains to be determined.

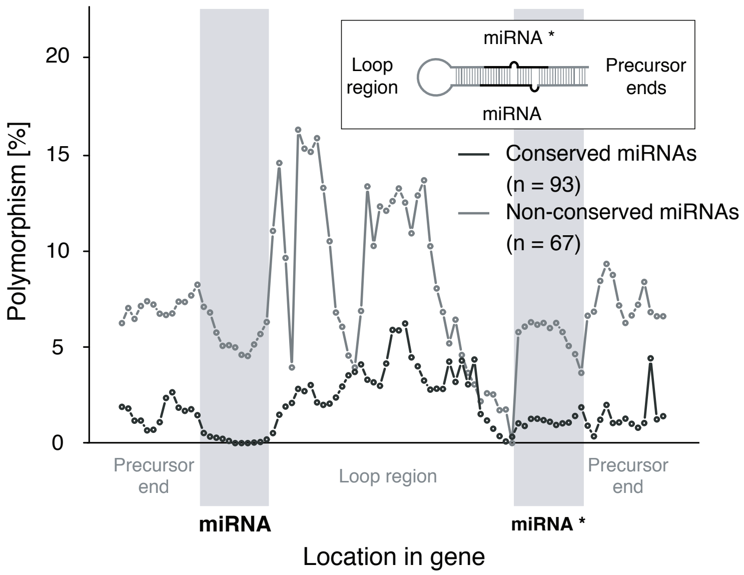


Figure 3.12: Polymorphism averaged over conserved and non-conserved miRNA genes by location in the stem loop structure (inset and as labeled at bottom). To facilitate visualization, lengths of the stem-loops were scaled relative to each other as described in Section 2.5.14.

3.1.6 Data Release

The PR prediction data set is available for download from The *Arabidopsis* Information Resource (TAIR) [178], as are fasta files for all accessions and annotated genes in which PRs are indicated.[2] Furthermore, PRs are visualized in Generic Genome Browsers [173] at TAIR[3] and at POLYMORPH.[4]

[2] ftp://ftp.arabidopsis.org/Polymorphisms/Polymorphic_Region_Predictions_Zeller_2008/

[3] http://gbrowse.arabidopsis.org/cgi-bin/gbrowse/arabidopsis/

[4] http://gbrowse.weigelworld.org/cgi-bin/gbrowse/polymorph/

3.1.7 Utility of Predictions for Functional Studies

Our predictions are immediately useful for functional studies in *Arabidopsis*. 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 *Arabidopsis thaliana* mutant collections [2]. 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.[5] Deletions or highly polymorphic sequences have been shown to underlie diverse phenotypes that segregate in *Arabidopsis thaliana* populations [e.g., 78], 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 [90]. 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.

3.1.8 Application of Our Methods to Other Data and Broader Relevance

Although mPPR was tailored for predicting polymorphic regions with *Arabidopsis* resequencing array data, it should be readily applicable to other resequencing array data sets with some modifications. In previous experiments with human, mouse and rice [e.g., 54, 68, 111], DNA hybridized to arrays was generated by pooling long-range PCR amplicons of selected regions. For *Arabidopsis*, the entire genomic DNA was subjected to isothermal amplification [29]. 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 [53, 54, 112], and sample-matched data sets that could potentially be used for training have been reported for mouse [e.g., 121] and have been generated for rice [111]. In rice, polymorphic regions have recently been identified by Regina Bohnert using an extended version of mPPR [10].

[5] `http://www.inra.fr/internet/Produits/vast/RILs.htm`

3.2 Transcript Normalization of Tiling Array Data

In this section, we present a novel normalization technique for tiling array data which aims to alleviate the effect of divergent sequence properties of oligonucleotide probes on hybridization intensity (Fig. 3.13 A). It was specifically designed to reduce the variability among probes complementary to the same transcript. Ideally, all such (perfect match) probes are expected to produce the same signal intensity because they all have approximately the same amount of target molecules bound. However, real tiling array data are characterized by a high degree of signal variability, even for probes measuring the same transcript [144]. Assuming that the *de facto* deviations of individual probes from ideal transcript intensities are largely due to sequence-dependent differences in affinity, our goal was to learn a regression function which estimates the deviation between the observed intensities of individual probes and the *transcript intensity* taking probe sequences as input. Additionally, we modeled a dependency on the observed hybridization intensity by employing an array of regression functions instead of a single one. Training each of these functions on a specific quantile-range of intensities yielded separate predictors for low, medium and highly expressed transcripts. Although conceptually simple, this quantilization model offers a high degree of flexibility as it does not make any assumptions on how probe sequence effects scale with transcript intensity, and can thus easily accommodate, e.g., saturation effects. For solving the regression problems, we used two different techniques: Support Vector Regression (SVR) and Ridge Regression (RR).

Applying this so-called transcript normalization (TN) technique to *Arabidopsis* tiling array data, we were able to demonstrate its effectiveness in reducing sequence bias and its utility as a preprocessing routine for transcript mapping. Transcript normalization has therefore become a pivotal preprocessing step for the identification of novel genes from tiling array data. All results presented here as well as in Section 3.3 are based on joint work with Stefan R. Henz, Sascha Laubinger, Detlef Weigel and Gunnar Rätsch (see p. 137 for author contributions) [204]. Computational experiments were conducted on data from a single sample (T_003, see Table 4.4).

3.2.1 Computational Experiments

The computational experiments detailed below were performed with the aim of (i) characterizing the probe sequence-bias in detail; (ii) showing that we can alleviate this sequence effect with our transcript-normalization method to an extent comparable to the previously published Sequence Quantile Normalization (SQN) method [145] and (iii) demonstrating that, unlike SQN and other normalization techniques evaluated in Munch et al. [120], Royce et al. [145], our transcript normalization improves the separation between exon hybridization and background signal.

After preprocessing and normalizing the array data to reduce inter-chip variability (as described in Zeller et al. [204], see Section 2.3.2), we partitioned the *Arabidopsis thaliana* genome into ≈ 300 regions while avoiding splits in annotated genes. Mapping perfect match (PM) probes to genome locations resulted in $\approx 10{,}000$ probes per region. We randomly chose 40% of these regions for training, 20% for hyper-parameter tuning and the remaining 40% as a test set for performance assessment (the test regions were further used for segmentation experiments in Section 3.3).

3.2.2 Alleviation of GC Bias

Hybridization intensity was found to be strongly correlated with differences in GC content of tiling probes. Extreme differences in GC content were associated with more than 4-fold changes in median intensity (Fig. 3.13). This probe sequence effect was reduced by all methods considered here. However, in part this effect can also be attributed to GC-richness of coding regions (Fig. 3.13; [73]). Position-specific sequence effects were further investigated with so-called quantile plots [145]. By analyzing the 90th intensity percentile, which is expected to be enriched for measurements of highly expressed transcripts, we focused this comparison on specific binding signals rather than on unspecific binding effects. The strongest reduction of mono-nucleotide sequence effects was clearly achieved with SQN, although positional sequence effects were reduced by all normalization methods considered here (Fig. 3.14).

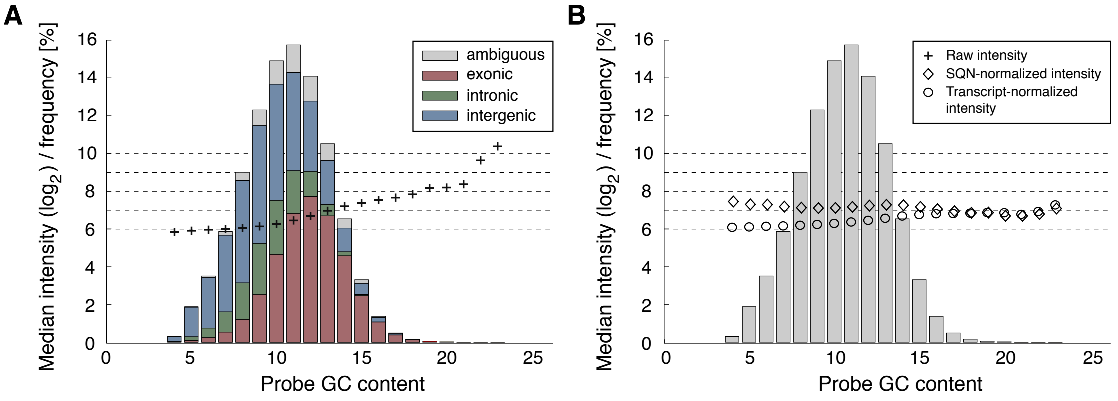


Figure 3.13: Median hybridization intensity as a function of probe GC content before and after normalization. **(A)** Dependency of median hybridization intensity on GC content of oligonucleotide probes. The histogram generated by partitioning probes according to the number of Gs and Cs in the 25mer probe is shown as a bar plot. The frequency of exonic, intronic and intergenic probes in each bin is color-coded (see inset). Median log-intensity per bin is shown before normalization (black crosses). **(B)** Median log-intensity as a function of probe GC content after the application of normalization methods (see inset; RR and SVR yielded virtually the same results, and therefore only the curve for RR is shown [204]).

3.2.3 Reduction of Transcript Intensity Variability

We next assessed transcript variability, i.e., to which extent individual probe intensities y_i deviate from the constant transcript or background intensity $\overline{y}_i$, for different normalization methods. In principle, transcript intensities $\overline{y}_i$ are unknown, but by using a robust summary statistics such as the median, we expected to obtain reasonable estimates. To assess

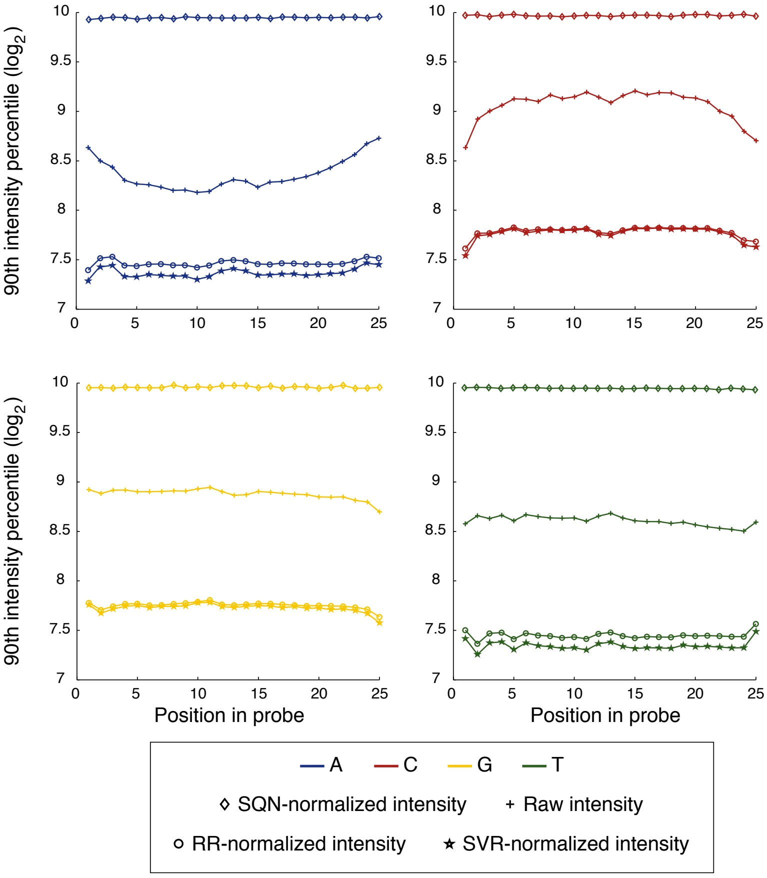


Figure 3.14: Position-specific quantile plots [144] for raw and normalized intensities (see inset). The effects of different nucleotides are displayed on different panels; the position within the probe sequence is indicated on the x-axis and the 90th intensity percentile on the y-axis.

transcript variability, we introduced two metrics, T_1 and T_2. Both relate the variability of normalized intensities $y_i - f(\boldsymbol{x}_i, y_i)$ to the variability of raw intensities, and values smaller than 1 indicate a reduction. We defined

$$T_1 := \frac{\sum_i |y_i - f(\boldsymbol{x}_i, y_i) - \overline{y}_i|}{\sum_i |y_i - \overline{y}_i|}$$

as the normalized absolute transcript variability and

$$T_2 := \frac{\sum_i (y_i - f(\boldsymbol{x}_i, y_i) - \overline{y}_i)^2}{\sum_i (y_i - \overline{y}_i)^2}$$

as the normalized squared transcript variability. SVR minimizes the so-called ϵ-insensitive loss closely related to the absolute error. In contrast, Ridge regression minimizes the squared loss. It was therefore not unexpected to observe smaller T_1 values for SVR and smaller T_2 values for RR (see Table 3.6). Both methods were found to effectively reduce transcript variability to approximately half the values of raw intensities. For SQN, however, we observed both T_1 and T_2 greater than 1 indicating increased transcript variability. One may argue that SQN is therefore not well-suited as a preprocessing routine for transcript mapping (see also Figs. 3.17 and 3.18). This likely reflects that TN normalization, in contrast to SQN, directly models and reduces the transcript variability.

Method	T_1	T_2
SQN	1.83	3.16
SVR	0.54	0.47
RR	0.58	0.44

Table 3.6: Within-gene variability after normalization compared between sequence-quantile normalization (SQN) and transcript normalization using either support vector regression (SVR) or Ridge regression (RR). For a definition of T_1 and T_2 see main text.

3.2.4 Modeling Transcript Amplification Bias Improves Normalization

In addition to probe sequence effects we found another major cause of intensity deviations from ideal constant transcript intensities. Intensity was observed to be generally higher near the 3' transcript end (Fig. 3.15). The most likely explanation for this is a bias in the T7-based linear amplification, which starts from an oligo-dT primer annealed to the polyA tail demarcating the 3' transcript end of the majority of plant mRNAs. The observed bias may be explained by an amplification process that terminates with a certain probability before the 5' transcript end is reached.

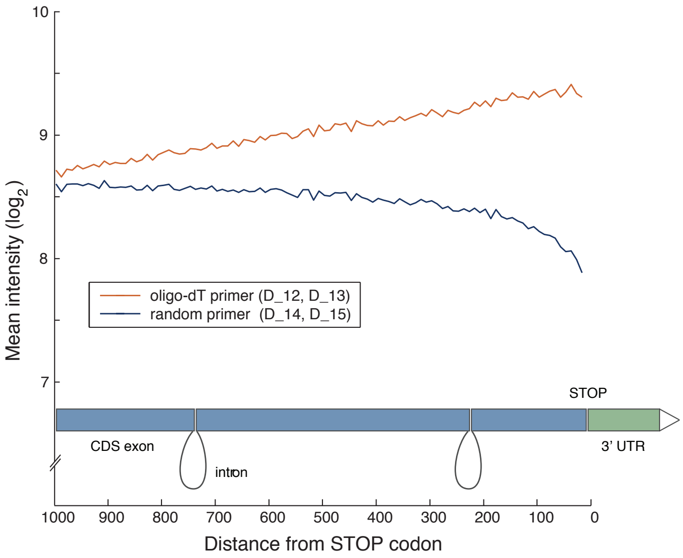


Figure 3.15: When oligo-dT primers were used for sample preparation, hybridization intensity on average decreased with increasing distance from the 3' transcript end. However, due to uncertainties in their annotation, instead of UTRs, STOP codons were considered for distance calculation here. We noted a different bias in data from arrays hybridized according to a protocol that involved random priming. Curves were generated with data from D12-D13 (oligo-dT) and D14-D15 (same tissues, but RNA amplified with random primers, see Table 4.4).

A correction for this bias can be simultaneously learned with sequence normalization by providing SVR or RR with additional features capturing the distance between a given probe and the 3' transcript end on the level of the spliced mRNA. With appropriate features, a piecewise-linear function can be estimated to quantify the distance effect on deviation from the transcript intensity. However, such features can only be employed for transcript normalization if transcript ends are known *a priori*. Although this may be the case when one is interested in identifying splice forms given the transcript start and end, this extension is of limited use for the identification of new transcripts and therefore not generally applicable. Nevertheless, we conducted experiments to quantify the 3' amplification bias relative to probe sequence effects (Fig. 3.15). When distance features were provided for training and testing, we obtained $T_1 = 0.44$, $T_2 = 0.33$ for SVR and $T_1 = 0.46$, $T_2 = 0.31$ for RR — a relative improvement of $\approx 25\%$ relative to transcript normalization without distance features.

3.2.5 Distinguishing Between Exon and Background Signal

In a naive approach to identify transcriptionally active regions (i.e., expressed exons) we used a simple threshold model on the hybridization measurements. Probes with intensities above the threshold were classified as exonic, ones below the threshold as untranscribed or intronic. Comparing the resulting classification of probe signals with the TAIR7 genome annotation [178], we computed precision and recall for a range of different threshold values. These are defined as the proportion of probes mapped to exons among all probes with intensities greater than the threshold value and the proportion of probes with intensities greater than the threshold value among all probes that are annotated as exonic, respectively.

In the following evaluation, we compared TN to SQN and additionally included two more normalization methods commonly applied to tiling array data. The first one makes use of so-called mismatch (MM) probes synthesized on most Affymetrix DNA microarrays. For each PM probe on the array, there is one MM probe that differs only at the central nucleotide. This design is intended to allow quantification of unspecific binding, and, theoretically, the specific binding component of the hybridization signal can be obtained by subtracting the MM intensity from the corresponding PM intensity[6] (designated PM-MM in the following). This strategy was evaluated in Royce et al. [145], and Munch et al. [120] proposed to use it as signal transformation prior to transcript identification with Hidden Markov Models. The second normalization strategy (denoted NM) is inspired by Naef and Magnasco [122] and implemented by fitting a linear model to position-specific contributions of different nucleotides to the overall PM probe signal [120, 145]. It can thus be seen as a simpler precursor method of our transcript normalization with three important differences: First, NM does not consider di- and tri-nucleotide contributions. Second, it attempts to directly relate probe sequence effects to the hybridization signal instead of modeling the deviation from transcript intensities. Finally, it fits all data together, instead of modeling intensity quantiles separately. Normalized intensities are eventually obtained by subtracting the predicted signal from the observed PM intensity.

For all normalization methods compared, we first verified that the distribution of exonic probe signals indeed exhibited only a heavy right tail relative to the distribution of background signals. We therefore do not need to consider a statistical test which considers both tails of the distribution, but can instead directly apply the naive thresholding approach outlined above because it assesses the overrepresentation of exon probes among those with high intensities. (Fig. 3.16). Comparing intensity distributions already re-

[6] Since log-transformation of the PM - MM signals is impossible when the MM signal is stronger than the PM signal, negative PM - MM differences are first set to 1.

vealed that transcript normalization effectively reduced the variance of background probe intensities compared to exon probe intensities (Fig. 3.16 C). However, this is clearly not the case for SQN (Fig. 3.16 E). Observing to which extent the distribution of exon intensities overlaps with that of background signals, one can appreciate that exon recognition accuracy is limited not only for this naive method, but also for more elaborate techniques (Fig. 3.16).

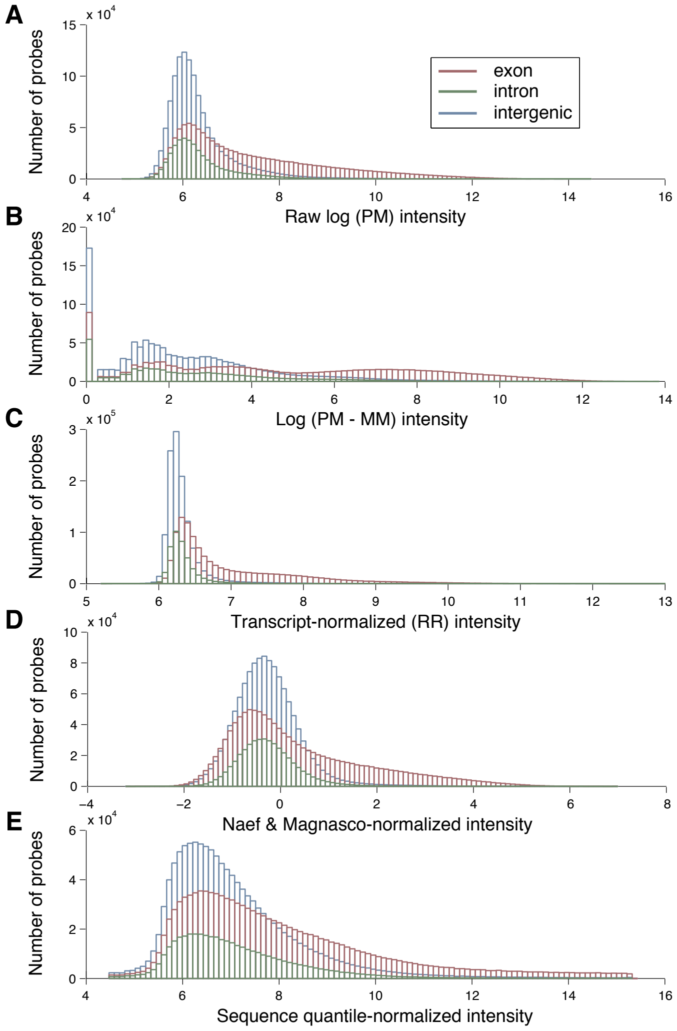


Figure 3.16: Intensity distributions for probes mapped to exons, introns and intergenic regions, respectively. The mapping to gene models was based on the TAIR7 annotation [178]. Probes partially mapped to more than one sequence type were excluded here. Intensity histograms are shown for raw PM intensities after $\log_2$ transformation **(A)**, for logged PM - MM intensities **(B)**, for transcript normalized intensities (only RR is shown here as SVR produces very similar distributions) **(C)**, for intensities normalized with the Naef & Magnasco method [120, 122, 145] **(D)**, and for intensities after sequence quantile normalization [145] **(E)**.

As a result of comparing different normalization methods with respect to their effect on the separation between exon and background signal (Fig. 3.17 A), we observed that the two transcript normalization methods SVR and RR yielded the greatest improvement compared to raw intensities. This was consistent across the whole range of possible trade-offs between precision and recall. In contrast, for SQN, PM-MM and NM, exon recognition deteriorated; most dramatically so for SQN (Fig. 3.17 A). However, when we sub-sampled the probes prior to thresholding and evaluation such that both, the exon and the background probe set had the same GC-content (as similarly done in Royce et al. [145]), the performance of these three methods recovered, and the largest improvement relative to raw intensities was observed for NM. Nevertheless even for the GC-corrected data set, the improvements obtained with SVR and RR are still greater than for all other methods evaluated (Fig. 3.17 B). One needs to consider, however, that this artificial subsampling strategy is of very limited use in practice, as it can not easily be applied to identify exon probes on a genome scale [58, 204].

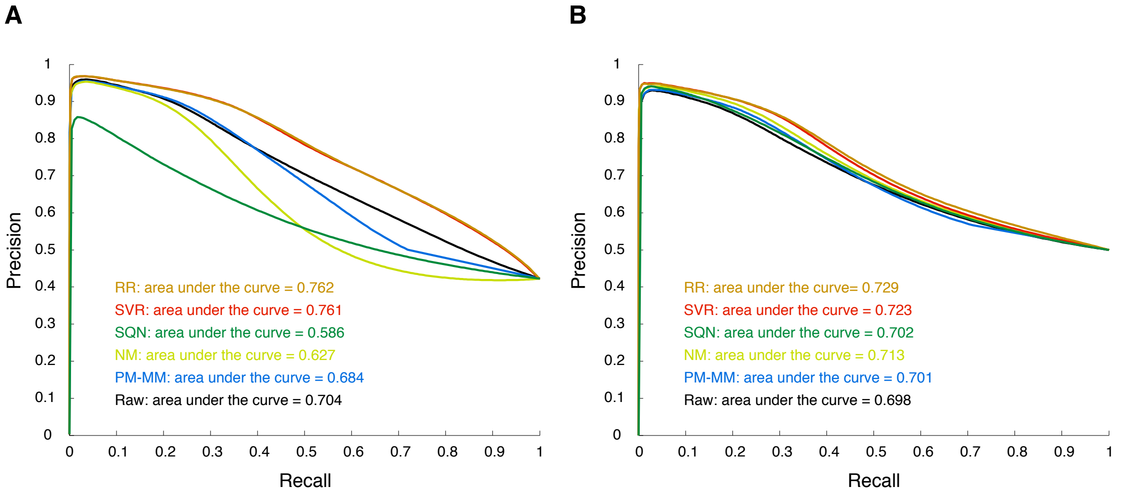


Figure 3.17: Separation between exon probe intensities and intensities of probes in regions annotated as untranscribed or intronic improves after normalization with SVR as well as after normalization with RR but not for the other methods evaluated here (PM-MM — intensity signal after subtraction of MM probe signal; NM — sequence normalization inspired by Naef and Magnasco [122] as implemented in [120, 145]; SQN — sequence quantile normalization [145]; SVR — transcript normalization using support vector regression; RR — transcript normalization using Ridge regression). **(A)** By varying the threshold value, we calculated the precision-recall curve from all probes in the test regions. **(B)** Prior to thresholding and precision-recall estimation, probes were sub-sampled to obtain the same GC-content among exonic and intronic / intergenic probes.

In a second experiment we only considered the transcribed regions of genes in the test set with the aim of distinguishing between exon and intron probe intensities. For this, we allowed a threshold to be chosen individually for each gene. Note that this problem is expected to be much easier than finding a single globally optimal threshold. However, the local thresholding approach cannot be directly applied when the transcript boundaries are not already known. For each gene we estimated the Receiver-Operator-Characteristic (ROC) curve separately and averaged them over all genes. We considered ROC curves instead of precision-recall curves (PRCs) here, since the class sizes vary among genes making PRCs incomparable. When comparing the area under the averaged ROC curves between genes with different expression levels (approximated by transcript intensity quantiles), we found that the ability to identify exons increases with increasing transcript intensities (Fig. 3.18). Again, we observed that the application of transcript normalization resulted in improved accuracy of exon probe recognition. Interestingly, transcript normalization performs consistently better than random guessing (area under the ROC curve of 0.5) even for the most weakly expressed genes. However, several of the other normalization methods considered here do not have this property. They probably introduce noise which severely complicates the detection of genes especially when these are only weakly expressed.

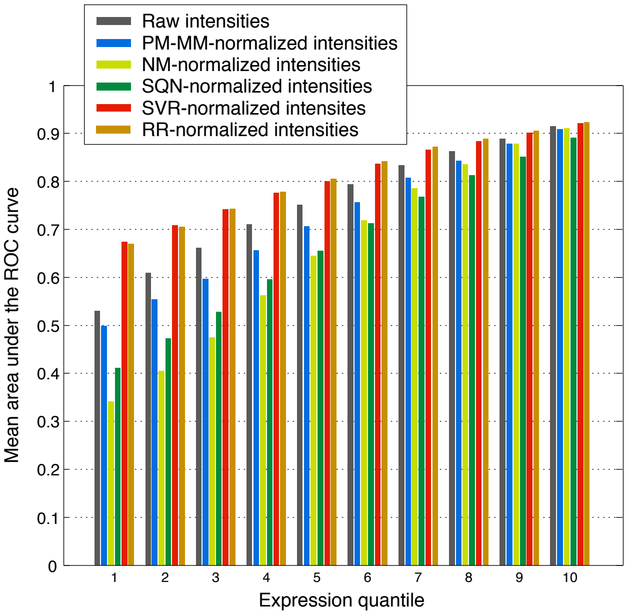


Figure 3.18: Separation between exon and intron probe intensities by gene expression and normalization methods (see inset). Gene expression quantiles were estimated as the median intensity of all corresponding exon probes (as annotated in TAIR7 [178]). For each gene the area under the ROC curve (auROC) was calculated from local thresholding, and for each gene expression quantile, auROC values were averaged over all genes in the same quantile.

To our surprise, these evaluations indicate that previously proposed normalization methods are ineffective as a preprocessing step for transcript identification from *Arabidopsis* tiling array data. This is particularly irritating because these methods were evaluated on human tiling array data confirming that exon recognition improved after normalization [120, 145]. However, Royce et al. [145] found only minimal improvements for a yeast tiling array data set [35] for SQN and NM. Although all these data sets were generated using Affymetrix tiling array technology and similar hybridization protocols, the investigated transcriptomes are very different in many aspects including gene density, genomic GC content, and relative differences in nucleotide composition between exons and intronic or intergenic sequences.

In the following, we examined the hypothesis that the discrepancy between evaluations for different organisms is largely due to differences in gene density. The smaller the proportion of expressed exons, the closer do we expect the overall intensity distribution to follow that of background probes. For human, where an estimated 1.5% percent of the genome are part of coding exons[7] [95], we conjecture that estimating sequence effects with a representative training sample may essentially amount to only modeling sequence effects for unspecific binding in the background with large error terms for the small proportion of exon probes. Consequently, the resulting sequence correction may be inappropriate for probes binding to expressed transcripts. Nonetheless, for a test sample containing very few expressed probes, such a normalization might still be beneficial with respect to reducing

[7] a slightly lower proportion is estimated for untranslated exons [95]

the number of false positive exon probe predictions. However, in the presence of a large proportion of exons as, e.g., for *Arabidopsis* (with about one third of its genome annotated as exonic), the proposed models are probably too simple to obtain good fits.

To test this hypothesis, we extended the NM normalization method in several steps into the direction of our transcript normalization approach and evaluated the effect by precision recall analysis as done before (Fig. 3.17). In the first experiment, we fitted the NM model exclusively on intergenic and intronic probes. Indeed this resulted in improved exon-background separation with an auPRC of 0.680 (compared to 0.627 originally obtained for NM). This supports the notion that the linear model might have problems fitting the possibly non-linear relationship between sequence effects and transcript abundance. Second, we substituted the original target variable for NM by the deviation from the corresponding transcript intensity, which is the regression target of our transcript normalization approach. This resulted in a thresholding performance of 0.684 as measured by the auPRC. In a third experiment, we used 20 independent NM models with the original NM regression target, each of them specifically trained and used to make predictions for one expression quantile. However, intensities normalized with the combined quantilized predictor exhibited extremely low thresholding accuracy of exon probe recognition (auPRC of 0.528). Surprisingly, this is however due to the fact that the quantilized predictor is now capable of also modeling exon intensities as NM normalization involves subtracting predicted from observed intensities. For a perfect predictor, such a normalization would eliminate intensity differences between exonic and background probes (assuming that we used sufficiently many quantiles) making exon recognition impossible. Thus, quantilization only makes sense in conjunction with the target variable used for TN, since in this setting normalization reduces the variance about ideal transcript intensities, but preserves intensity differences that are due to different quantities of bound target. Finally, to confirm this, we combined the last two extensions (quantilization and TN target variable) and obtained an auPRC of 0.752, which is a significant improvement over raw intensities and only slightly worse than for TN (RR) (0.762, Fig. 3.17 A). The remaining performance difference can probably be attributed to a richer set of features as well as regularization employed in TN.

Taken together, these results are consistent with our hypothesis that previously proposed normalization methods may be appropriate for correction probe-sequence effects of unspecific (background) binding. However, they do not model sequence effects sufficiently well across the whole intensity range to benefit exon recognition in organisms with compact, relatively gene-rich genomes. Although we have not addressed this directly, we conjecture that the accuracy gain of our transcript normalization method is not an organism-specific phenomenon. It remains to be tested, to which extent the more complex model would be beneficial for tiling array data from human samples.

3.3 Margin-Based Segmentation of Tiling Array Data (mSTAD)

For the *de novo* identification of transcribed regions from tiling array data, we developed mSTAD (margin-based segmentation of tiling array data). It is a supervised machine learning-based approach which is trained on regions around known genes using the discriminative HM-SVM framework and is subsequently able to accurately predict transcriptionally active regions genome-wide.

Conceptually, mSTAD constitutes a three-fold advancement over a previously proposed transcript mapping method originally applied to yeast tiling array data [72]. First, mSTAD is designed to recognize spliced transcripts and explicitly models introns. Second, a flexible noise model is fitted during training which does not make strong assumptions about the distribution of the hybridization signal. Third, supervised training and cross-validation procedures for adjusting hyperparamters bypass the need to manually tune internal parameters such as the expected number of transcripts which facilitates the application of mSTAD to larger eukaryotic genomes.

Before discussing its application to data sets of interest to biologist researchers, we present a thorough assessment of mSTAD's prediction accuracy compared to the naive thresholding approaches introduced in Section 3.2, as well as to more elaborate transcript mapping techniques used in practice. The results of this section were (partially) published in Zeller et al. [204] and generated together with Gunnar Rätsch, Timo Sachsenberg and Sascha Laubinger (see also p. 137).

3.3.1 Assessment of Segmentation Accuracy Comparing Different Normalization Techniques

For a first proof-of-concept experiment, we used triplicate array data from a single sample (T_003, see Table 4.4) and considered genomic regions around known genes. Each of these regions (from the test set described in Section 3.2) contained exactly one annotated gene, as well as up to 1 kbp of flanking intergenic sequence, truncated before the next known transcript was encountered. We randomly chose 100 of these for training, 100 for model selection and 500 other regions for evaluation. When comparing our method with the two simple thresholding approaches described in the previous section, one should bear in mind that the local thresholding method has an advantage because threshold values are optimized individually per gene. It cannot be directly applied to *de novo* detection of exon probes, since the threshold value is based on the expression levels of genes yet to be identified, and a partition into single-gene regions is not available unless all genes are known. In contrast to that, optimizing a global threshold can be realistically utilized for exon probe identification.

For comparative performance assessment we applied mSTAD and the two thresholding methods to raw as well as normalized hybridization intensities discussed in Section 3.2. mSTAD returned a segmentation of the tiling path into ("transcriptionally active") exons, introns and intergenic regions, whereas the thresholding methods partitioned tiling probes into "active" (expressed) ones corresponding to exons as well as "inactive" (background) probes. A comparison of the accuracies for exon probe recognition between the three methods (Table 3.7) revealed that our method works considerably better than global thresholding, and even slightly better than local thresholding. Moreover, we could reconfirm the findings of the previous section that transcript normalization significantly improved discrimination between transcriptionally active and inactive regions not only when thresholding on a per-probe basis is applied, but also with a considerably more complex segmentation algorithm.

	Thresholding		mSTAD
	global	local	
Raw intensities	71.1%	79.3%	79.7%
Sequence quantile normalization	66.1%	75.3%	73.5%
Support Vector Regression	73.9%	82.0%	83.0%
Ridge Regression	73.8%	82.0%	83.8%

Table 3.7: Accuracy of transcript identification given regions with exactly one gene. Accuracy is defined as the sum of true positive and true negative exon probes over the total number of probes in a gene. Evaluation is based on data from inflorescence tissue (T_003, see Table 4.4).

3.3.2 Prediction Accuracy in Comparison to Other Transcript Mapping Methods

Next, we performed computational experiments to compare the performance of mSTAD (trained discriminatively with the HM-SVM algorithm and thus called “mSTAD HM-SVM” in the following) to other methods commonly applied for the detection of transcriptionally active regions (TARs).

Among previously proposed transcript mapping methods, the so-called transfrag method, originally developed for the analysis of tiling array data from human samples [85], has become a very popular tool [28, 65, 83, 108, and others]. It is based on determining expressed probes in a local context by statistical testing. On top of that, a sliding window approach is taken for the identification of TARs. It searches for regions with a certain number of consecutive positive probes (minRun parameter) which are interrupted by no more than a certain number of negative probes (maxGap parameter). In addition to those, a few more hyperparameters are to be specified by the user. However, their influence on prediction accuracy is not directly apparent and an optimal hyperparameter set is difficult to determine *a priori*.

Hidden Markov Models (HMMs) have also been devised for the analysis of tiling array data, at first with the purpose of identifying transcription factor binding sites in ChIP-chip experiments [42, 77, 104], but later also for transcript mapping [77, 120]. They offer a principled alternative to sliding window methods with the benefit that they require less manual tuning. As Hidden Markov Support Vector Machines (HM-SVMs) are a label sequence learning technique which is closely related to HMMs, we could easily adopt an HMM training algorithm for mSTAD while leaving the state model and decoding algorithm unchanged (in the following abbreviated "mSTAD HMM"). However when trained on a representative set of annotated genes regardless of their expression state, HMM predictions were found to be very unspecific. Precision could be improved — albeit at the cost of a lower recall rate — when training examples were selected around genes with expression above a certain level.

To minimize the effect of unreliable probe annotation on the assessment of prediction accuracy, we evaluated all methods on a set of 1000 genes annotated in TAIR7 [178], for which the complete structure was confirmed by full-length cDNA sequences. All methods were trained (if necessary) on triplicate tiling array data from root tissue (D_001, see Table 4.4) on examples sampled from regions disjoint from the test set.

Prediction accuracy was compared by means of precision-recall analysis on the level of individual probes, exons, and introns; furthermore we assessed how accurately exon boundaries were predicted with respect to the resolution of the tiling array (Fig. 3.19, also for definitions of precision and recall). To facilitate a meaningful comparison, various prediction sets with different trade-offs between precision and recall were generated for each method. The transfrag method was evaluated for 900 different hyperparameter combinations corresponding to a wide range of precision-recall values. For mSTAD HMM, a number of different training sets were prepared containing genes with increasing expression level resulting in increasing precision of the corresponding predictions. For mSTAD HM-SVM we manipulated transition scores after training to adjust the trade-off between precision and recall (as indicated in Fig. 3.19 and detailed in Section 2.6.5).

As a result of the method comparison, we found that all other methods evaluated are less accurate than mSTAD trained with the HM-SVM algorithm in terms of probe-level accuracy (Fig. 3.19 A). Only when restricted to highly expressed genes does the HMM training algorithm achieve precision and recall comparable to the HM-SVM method. The transfrag method is generally less accurate than mSTAD. When considering correctly predicted exons, we found that mSTAD HMM was more accurate for some training sets with low gene expression threshold, but in the high-precision regime, the HM-SVM training yielded slightly higher accuracy (Fig. 3.19 B). High exon recall of the HMM predictions was, however, largely the result of comparably low recall for introns. In cases of introns missed by mSTAD HMM, a large exon prediction typically spanned several annotated exons and the intervening introns. This behavior positively influences exon accuracy, but negatively impacts intron and per-probe accuracy (which was found to be up to ten percentage points lower than for comparable HM-SVM instances in terms of the average of precision and recall). Conversely, mSTAD HM-SVM was superior — particularly more sensitive — when compared to mSTAD HMM on the intron level (Fig. 3.19 C). Furthermore, exon recognition accuracy is traded off more flexibly against intron accuracy by mSTAD HM-SVM compared to mSTAD HMM (Fig. 4.7). When assessing how accurately exon boundaries could be predicted (with respect to the tiling probe resolution), we found that mSTAD HM-SVM was roughly as accurate as mSTAD HMM: the average of precision

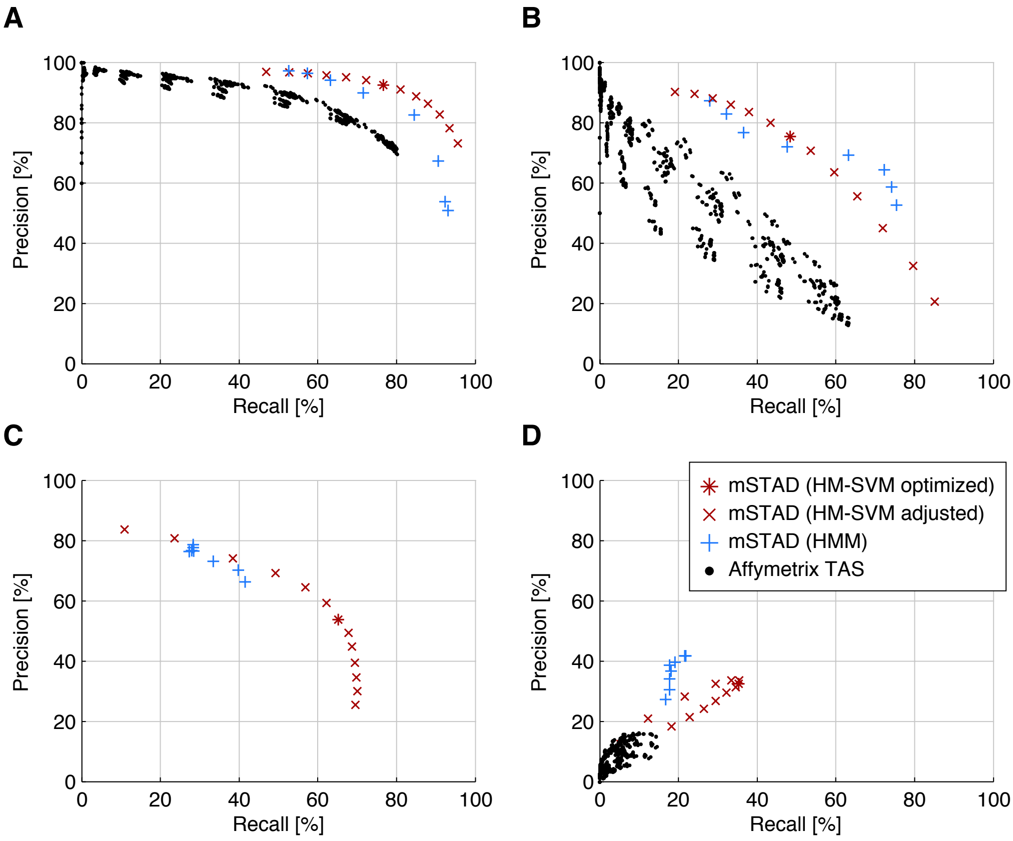


Figure 3.19: Precision-recall curves showing mSTAD's performance in comparison to other methods. The HM-SVM approach is compared to an equivalent HMM trained generatively and to the transfrag method (implemented in the Affymetrix TAS package). To trade off precision and recall (which is the same as sensitivity), transition-scores of the HM-SVM approach were manipulated after training, different gene sets were used to retrain the HMM, and for the transfrag method hyperparameters were varied in a grid search (see inset and main text for details). **(A)** Probe-level evaluation. Recall is defined as the proportion of exonic probes contained in predicted exons relative to all annotated exon probes. Precision indicates how many predicted exon probes are annotated as such. **(B)** Evaluation of predicted exons. Here, recall is defined as the proportion of annotated exons for which all included probes are predicted as such relative to all annotated exons. Precision is equal to the proportion of predicted exons for which all probes also belong to annotated exons. **(C)** Evaluation of intron predictions. Precision and recall are defined as for exon predictions but with respect to introns and intronic probes. Because it only discriminates between exonic and intergenic probes, the transfrag method could not be evaluated in this category. **(D)** Evaluation of exon-boundary predictions. Here, only those exon predictions that include all annotated exon probes but none of the surrounding probes annotated as intronic or intergenic are treated as correct predictions.

and recall was 34% and 32% for the best discriminative and generative model, respectively — roughly twice as accurate as the transfrag method (Fig. 3.19 D).

Owing to its convincing accuracy justifying the computationally more demanding HM-SVM training, we decided to use mSTAD HM-SVM for annotating transcriptional activity in the *Arabidopsis thaliana* genome.

3.3.3 Genome-Wide Analysis of the Arabidopsis Transcriptome in Various Tissues and Developmental Stages

Together with Sascha Laubinger, Stefan R. Henz, Timo Sachsenberg, Christian K. Widmer, Naira Naouar, Marnik Vuylsteke, Bernhard Schölkopf, Gunnar Rätsch and Detlef Weigel we analyzed tiling array data from diverse tissues and developmental stages of *Arabidopsis thaliana* (D series, see Table 4.4) with the goal to discover new transcripts not present in the current genome annotation (see p. 137 for author contributions) [97]. For genome-wide predictions of TARs, we applied mSTAD with a statemodel that approximated continuous gene expression values by 10 discrete levels (Fig. 3.20).

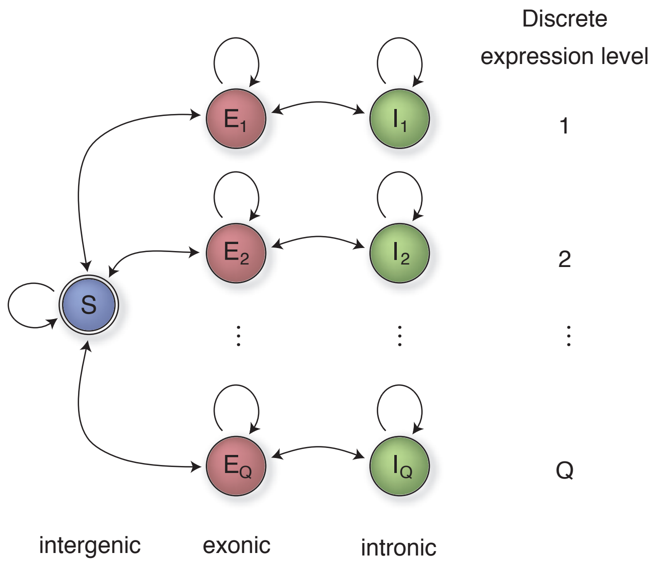


Figure 3.20: Simplified state model employed in mSTAD. For each of $Q = 10$ discrete expression levels, there is a submodel with exon and intron states. Modeling introns depending on the expression level of the surrounding exons allows to carry expression information along the whole transcript (see also Fig. 2.5 for the complete state model).

New TARs in the Arabidopsis Genome

When comparing a genome-wide sample of all TARs predicted with mSTAD to annotated genes, we found that the predictions are generally accurate for the more highly expressed half of genes (Fig. 3.21). Accuracy was also found to be inversely correlated with repeat content which is higher in and around centromeres (Fig. 3.22). We therefore restricted further analyses to a selected set of high-confidence TARs (Fig. 3.23 A). These contained a minimum number of four consecutive probes predicted to be exonic, had predicted discrete expression level between 6 and 10, and at most 25% repetitive probes. When evaluated on the same set of 1,000 full-length cDNA-confirmed genes already used to compare different transcript mapping methods (Fig. 3.19), the recall rate of high-confidence exons was about half compared to all exons, as about one half of the exon predictions passed the above filters. Precision improved from 75.3% to 78.9% on the exon level and from 90.8% to 98.9% on the exon overlap level.

Genome-wide, high-confidence TARs make up 37 to 50% of the total length of all predictions depending on the tissue analyzed. More than 97% of these high-confidence TARs overlap at least 25 bp with annotated exons (Fig. 3.23). Between 26 and 36% of the remainder overlap with cDNAs and ESTs but not with annotated transcripts. In summary, there are between 1,107 and 1,947 predicted high-confidence TARs per sample, for a total length of 242 to 406 kb, that are neither included in the current annotation nor covered by sequenced cDNA or EST clones. Non-redundantly across all analyzed tissues, unannotated high-confidence TARs cover 2,127 kb, and about 46 kb of the *Arabidopsis thaliana* genome are detectable with high confidence as transcriptionally active in all analyzed tissues despite their intergenic annotation. The observed difference between the union and the intersection of unannotated high-confidence TARs predicted for different samples likely reflects large changes in gene expression between different plant tissues and developmental stages.

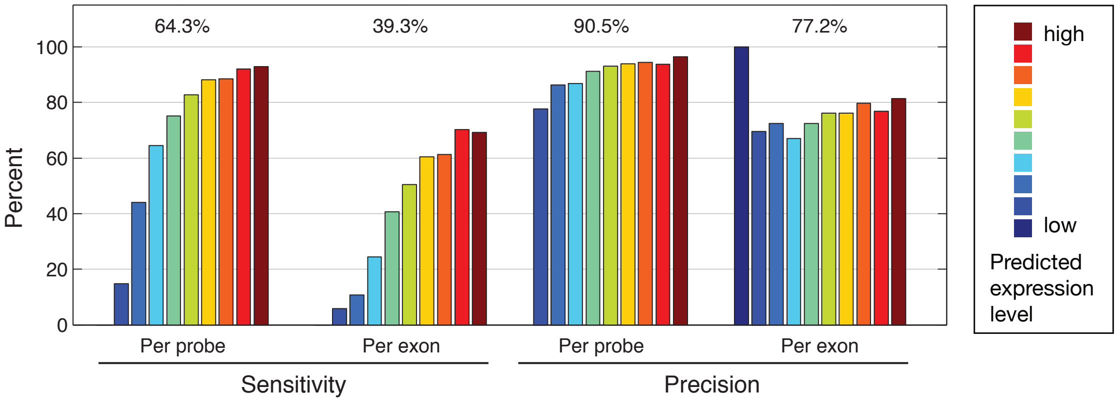


Figure 3.21: Segmentation accuracy for root tissue (D_001, see Table 4.4) across ten discrete expression levels (see inset). On the probe-level, sensitivity (recall) is defined as the proportion of exonic probes contained in predicted exons relative to all annotated exon probes. Precision indicates how many predicted exonic probes are annotated as such. On the exon level, we define sensitivity (recall) as the proportion of annotated exons for which all contained probes are also predicted to be exonic. Here, precision is defined as the proportion of predicted exons which do not contain any probe annotated as intronic or intergenic.

Among the unannotated high-confidence TARs, 14 to 31% are specifically detected in a single sample, with inflorescences and senescing leaves showing the highest proportion (Fig. 3.23 B). Whether these predictions indeed correspond to expressed transcripts was tested for some of them by reverse transcriptase followed by PCR (RT-PCR). From TARs which do not overlap with known cDNAs or ESTs, a subset of 47 segments was selected so that different lengths as well as different predicted expression levels were covered. We could confirm by RT-PCR more than three quarters (37) of these 47 predicted segments as transcribed (see Fig. 3.23 C for several examples).

The "Dark Matter" of the Arabidopsis Genome

The fact that there is not only a good correspondence between TARs and annotated genes, but that we also achieved high success rates for RT-PCR validation of new TARs, corroborates mSTAD's high prediction accuracy. Interestingly, despite a rather complete

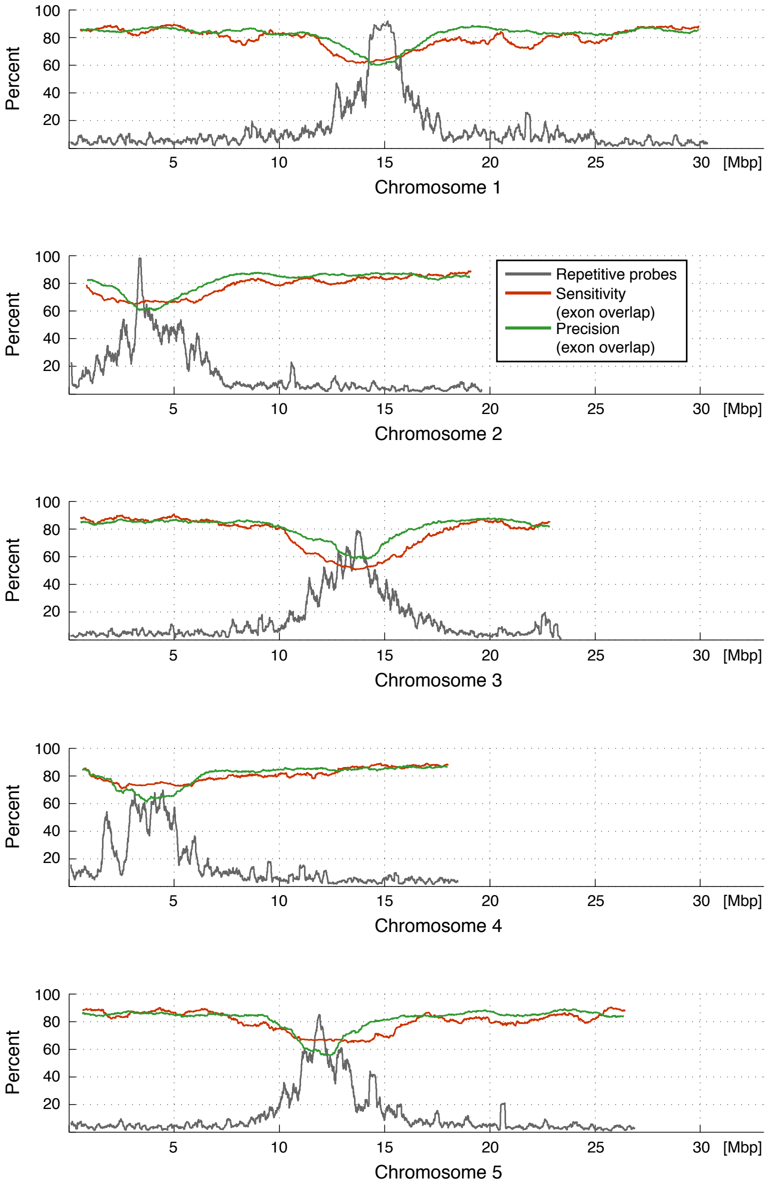


Figure 3.22: Congruence between genes annotated in TAIR7 [178] and transcriptionally active regions (TARs) predicted by mSTAD. On exon overlap level, sensitivity (same as recall) is defined as the proportion of annotated exons containing at least one probe predicted to be exonic relative to all annotated exons. Precision is defined as the proportion of predicted exons overlapping by at least one probe length (25 nt) with an annotated exon. Both measures were plotted in a sliding window across 2,000 exons along the five *Arabidopsis thaliana* chromosomes together with information on repetitive probes (window of 5,000 probes; see inset).

genome annotation for *Arabidopsis thaliana* that is based on extensive cDNA cloning and previous use of tiling arrays [e.g., 198], we could detect more than one thousand additional TARs per sample analyzed.

Initial tiling array-based transcriptome studies of the human genome reported evidence of transcription from much more genomic sequence than was annotated as protein coding loci [e.g., 86]. This discrepancy between annotated genes and tiling array-based evidence for transcription, estimated to be as large as one order of magnitude [86], prompted people to speculate about the "dark matter" of the (human) genome [79]. To bridge

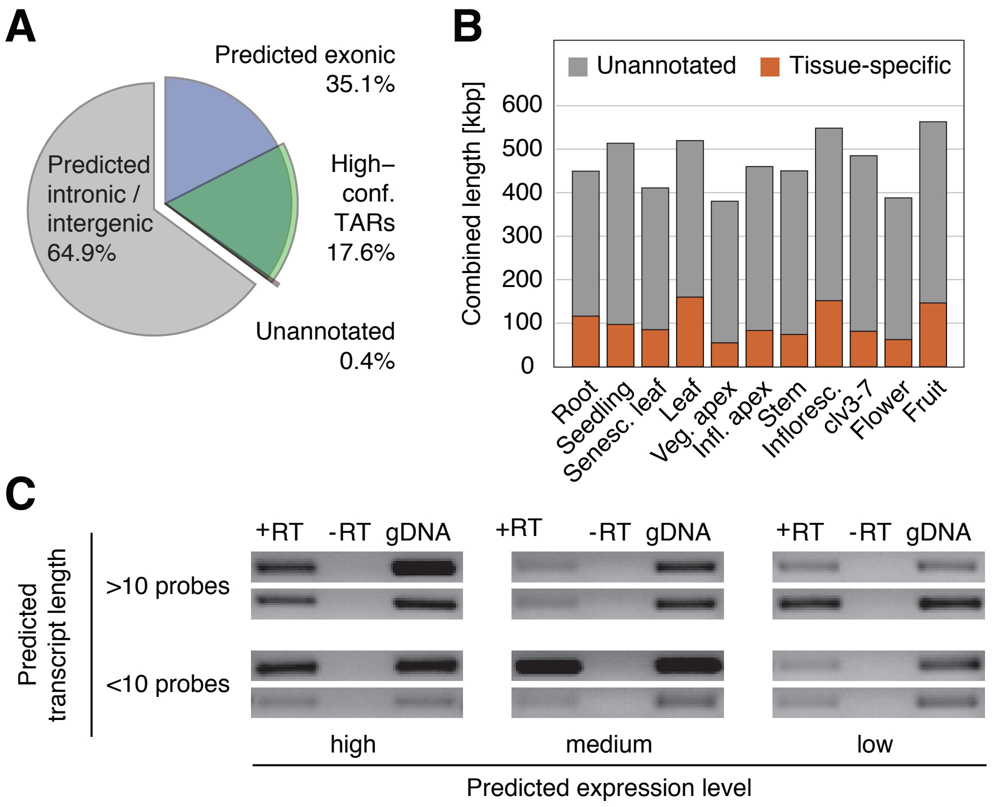


Figure 3.23: **(A)** Proportion of all transcriptionally active regions (TARs), high-confidence TARs (see main text for definition), and unannotated TARs (high-confidence predictions that do not overlap with any annotated exon by at least 25 base pairs). Percentages are based on combined length of each class. **(B)** Proportion of sample-specific TARs among all unannotated high-confidence TARs. **(C)** Examples of RT-PCR validation of predicted new transcripts.

this gap, research has more recently focused on characterizing genes that produce RNA capable of performing cellular functions rather than being translated into protein [65, 87, 147, 208, among many others]. Despite our finding of unannotated transcripts, the proportion of annotated transcripts among TARs predicted from tiling arrays appears to be much higher in *Arabidopsis* permitting the conclusion that the *Arabidopsis* annotation is relatively complete. Nevertheless, tiling array analysis of *Arabidopsis* mutants impaired in DNA methylation or RNA quality control has revealed over two hundred non-coding transcripts that are normally transcriptionally silenced, indicating that the *Arabidopsis thaliana* genome has at least the potential to generate a large number of transcripts from intergenic regions [26, 93, 208].

The Non-Polyadenylated Arabidopsis Transcriptome

Previous analyses with whole-genome tiling arrays have focused on the polyadenylated portion of the *Arabidopsis* transcriptome [26, 175, 198, 208]. However, studies in several other organisms have suggested that there is a large fraction of non-polyadenylated RNAs [e.g., 28, 65]. In order to revisit this question in *Arabidopsis*, we isolated total RNA from two different tissues, whole seedlings and inflorescences, and used two different experimental protocols to generate tiling array hybridization data. RNA was prepared for reverse transcription with either an oligo-dT primer, which targets only polyadenylated RNA (polyA(+)), or random primers, which target both RNAs with and without a polyA tail (polyA(+/-)). We then applied the mSTAD algorithm to these data to detect transcription from unannotated regions. When we subtracted high-confidence TARs found in at least one polyA(+) sample from the TARs found in both polyA(+/-) samples, TARs totaling less than 100 kb were identified as potential polyA(-) transcripts (Fig. 3.24 A). These regions represent less than 0.1% of the entire genome, which appears to be very low compared to results reported for *Caenorhabditis elegans* tiling array studies using the transfrag method [65]. To rule out the possibility that this discrepancy is a computational artifact, we applied the transfrag method also to our tiling array data [85]. This method led to similar estimates of non-repetitive polyA(+/-) specific transcribed fragments (transfrags), with a combined length of about 250 kb, or 0.2% of the genome (Fig. 3.24 B).

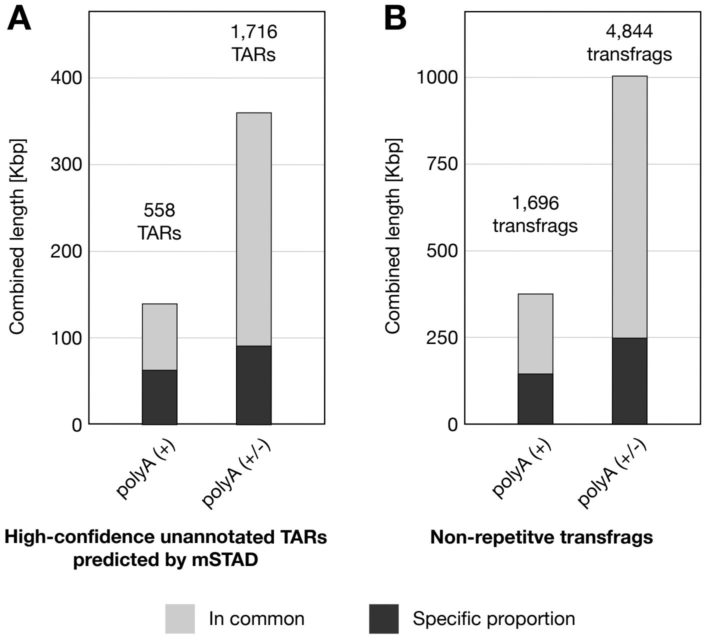


Figure 3.24: Non-polyadenylated transcripts. Proportion of unannotated transcripts found in common or exclusively in either polyA(+) samples and polyA(+/-) samples, respectively, as determined with two independent methods. **(A)** TARs predicted with mSTAD **(B)** Transfrags predicted using Affymetrix software [85].

These results imply that non-polyadenylated transcripts are much less abundant in *Arabidopsis* compared to human and *Caenorhabditis elegans*. For these organisms, tiling array-based studies indicated that about half of all transcripts are not polyadenylated [28, 65]. It is already known that specific classes of plant transcripts are generated in a different manner than in animals. For example, some human microRNA precursors are transcribed by RNA polymerase III and hence are not polyadenylated, while *Arabidopsis* miRNA precursors feature characteristics of RNA polymerase II-generated transcripts [13, 196]. Another reason might be differences in 3' end processing. For example, histone mRNAs in land plants are polyadenylated, which is in contrast to histone mRNAs in animals that are subject to a unique form of 3' end processing resulting in a hairpin protecting the 3' end from RNA degrading enzymes [22, 23, 24, 41, 193].

TARs Predicted by mSTAD are Part of the At-TAX Community Resource

We made our results easily accessible to the research community with the introduction of the so-called At-TAX (***A****rabidopsis* ***t****haliana* **T**iling **A**rray **Ex**pression Atlas) online resource.[8] It includes a customized Generic Genome Browser [173] that displays TARs predicted by mSTAD across the genome as well as all raw expression values for each probe in all analyzed samples (see Fig. 3.25 for an example).

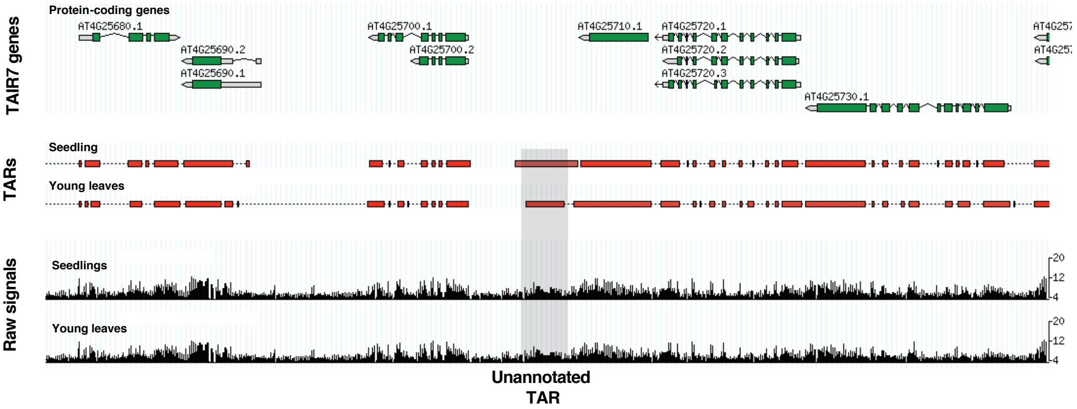


Figure 3.25: A customized Generic Genome Browser is part of the At-TAX online resource. The screenshot displays predicted TARs (middle) and raw hybridization signals (bottom) along the chromosome.

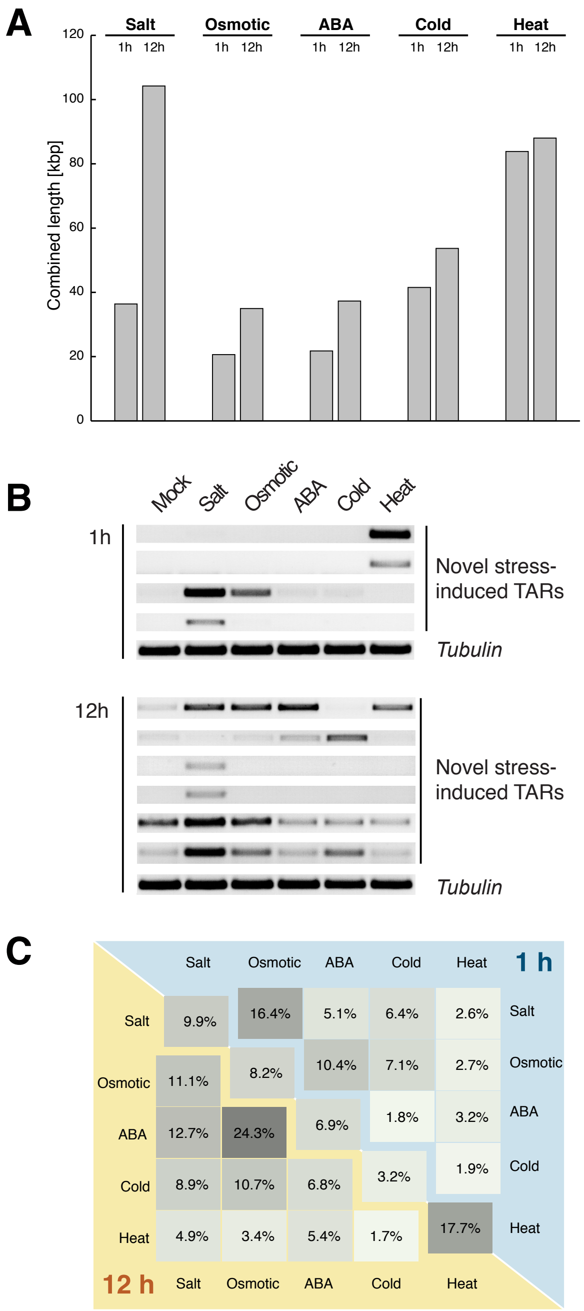
3.3.4 Identification of New Stress-Induced Arabidopsis Transcripts

After we had analyzed the *Arabidopsis* transcriptome of different plant organs and developmental stages (D series, see Table 4.4) to identify transcriptionally active regions (TARs) genome-wide, we investigated the plant transcriptome under various abiotic stress conditions (S series, see Table 4.4). To this end, we extended our previously developed methods by an additional statistical test for the detection of stress-induced TARs. The following results are based on joint work with Stefan R. Henz, Christian K. Widmer, Timo Sachsenberg, Gunnar Rätsch, Detlef Weigel and Sascha Laubinger (see p. 137 for author contributions) [205].

Identification of Stress-Responsive, Unannotated TARs

In a first step we applied mSTAD to discover new TARs outside of annotated exons as described above. New high-confidence TARs (see Section 3.3.3 for definition) were tested

[8] `http://www.weigelworld.org/resources/microarray/at-tax`



	Salt	Osmotic	ABA	Cold	Heat
Salt	9.9%	16.4%	5.1%	6.4%	2.6%
Osmotic	11.1%	8.2%	10.4%	7.1%	2.7%
ABA	12.7%	24.3%	6.9%	1.8%	3.2%
Cold	8.9%	10.7%	6.8%	3.2%	1.9%
Heat	4.9%	3.4%	5.4%	1.7%	17.7%

Figure 3.26: Identification of new, stress-induced transcripts. **(A)** Combined length of all unannotated high-confidence predictions of transcriptionally active regions (TARs) identified under various stress conditions and at two different time points is shown. **(B)** RT-PCR validation experiment of 10 new TARs confirming stress-induced expression. **(C)** Pairwise overlap of new TARs induced by different stresses at the same time point (upper and lower triangular matrix) or by the same stress at different time points (diagonal).

for significantly higher expression in stress-treated samples relative to mock controls using the Mann-Whitney-U test. Per individual stress treatment, we found 82 to 338 new high-confidence TARs for which hybridization signal was significantly higher under stress at the 5% level. These stress-induced TARs covered 21 kb to 104 kb of the genome (Fig. 3.26 A). The accuracy of this approach was further examined by reverse transcription PCR (RT-PCR) validation experiments. Indeed, in the tested cases TARs predicted to be strongly stress-responsive are more abundant in stress-samples than in the corresponding mock control (see Fig. 3.26 B for several examples).

The size of individual stress-induced TARs ranged from approximately 135 bp (4 tiling probes) to almost 2 kb (53 tiling probes). Most of them were found after 12 hour salt stress treatment, while the fewest were identified after one hour osmotic stress or ABA treatment (Fig. 3.26 A). We also asked how specific the stress response of new TARs is. In a pairwise comparison, we found the greatest overlap between new TARs after salt stress, osmotic stress and ABA treatment (Fig. 3.26 C), resembling the pattern obtained for annotated genes in a similar analysis [205]. However, the overall percentage of overlap was lower for stress-induced TARs than for annotated genes.

TARs Supported by Massively Parallel Signature Sequencing (MPSS)

As an independent experimental validation of tiling array-based predictions of transcriptional activity, data from massively parallel signature sequencing (MPSS) is ideally suited. Even before it has become feasible to assay transcriptional activity with short read sequencing technology [106, 110, 118, among others], collecting small sequence tags of approximately 15-20 nt from the ends of transcripts has been an established sequencing-based alternative to microarray technology for quantitative measurements of gene expression [17, 114, 115, 187].

To assess the fraction of TARs predicted by mSTAD that are also supported by MPSS, we compared TARs to 20-bp signatures that were collected from five *Arabidopsis* tissues and several mutants as well as one plant hormone treatment (salicylic acid applied to leave tissue). We only considered signatures that could be reliably mapped to the genome and were deemed significant in a previous analysis [114]. On average per sample, $\approx 42\%$ of high-confidence TARs were found to have a reliable and significant MPSS tag mapped within their boundaries when pooling MPSS tags that were originally mapped to either strand of the genome (Table 3.8). The median support for unannotated, stress-induced TARs was 19% and therefore significantly lower than for all high-confidence TARs. Nevertheless, the MPSS support for stress-induced TARs is comparable to that observed for annotated exons suggesting only a moderate increase in false discovery rate (Table 3.8).

Genomic Location and Conservation of New Transcripts

To characterize new stress-responsive TARs in more detail, we determined conservation of the genomic regions that give rise to these TARs in three other plant species for which complete or nearly-complete genome sequences are available, *Poplar trichocarpa*, *Oryza sativa* and *Sorghum bicolor* [5, 59, 185, 201]. Compared to annotated exons, new stress-specific TARs are in general much less conserved (Fig. 3.27 A). This finding could reflect that these new TARs are evolutionarily younger or less stable. Alternatively, if these TARs are mostly non-coding, primary sequence conservation might be less important. New stress-specific TARs in the genome might either constitute unannotated exons of known genes or they might be independent genes. A simple indicator for these alternatives can be the distance of new TARs to annotated genes. Per sample, we identified between 21 and 69 unannotated stress-specific TARs separated by more than 500 bp from the nearest annotated genes (Fig. 3.27 B, examples shown in Fig. 3.27 C; other samples see 4.9), while

Sample	high-confidence TARs		stress-induced TARs	
	Number of TARs	MPSS support	Number of TARs	MPSS support
S_001	58,740	42.4%		
S_002	61,536	40.9%		
S_003	57,840	42.7%		
S_004	61,398	40.2%	132	18.9%
S_005	56,180	42.2%	373	14.2%
S_006	58,471	41.3%	82	22.0%
S_007	53,368	42.9%	132	24.2%
S_008	57,431	41.1%	107	7.5%
S_009	52,568	42.7%	143	21.7%
S_010	60,209	41.8%	159	23.9%
S_011	56,054	43.1%	213	16.9%
S_012	56,729	41.6%	338	16.3%
S_013	54,790	42.4%	293	19.1%
annotated exons	158,070	25.2%	18,227	29.3%

Table 3.8: TARs predicted by mSTAD per sample and support by MPSS tags. For comparison, MPSS support for exons annotated in TAIR7 [178] is also shown; the last two entries in this row correspond to exons in annotated genes detected to be upregulated on tiling arrays for any of the analyzed stresses [205]. All numbers are based on reliable and significant MPSS tags from Meyers et al. [114].

others are in close proximity to or even abut annotated genes (examples in Fig. 3.27 D). Because our method did not identify the strand from which transcripts arise, we examined some of these cases by reverse transcription followed by PCR (RT-PCR). In one case, there is apparently an additional exon that is specifically induced under salt stress, but not others. Alternatively, this TAR might correspond to an independent stress-specific transcript with a large overlap to the annotated gene (Fig. 3.27 D, left). In another case, a minor transcript form is present under all conditions, but becomes more abundant under a specific stress (Fig. 3.27 D, middle). In a third case, it appears that a constitutive exon of a stress-responsive gene has simply been missed in previous annotation efforts (Fig. 3.27 D, right).

Incorporation of Stress Data into the At-TAX Online Resource

For the whole stress data set described here, we integrated single probe intensities as well as predicted TARs along the chromosomes into our *Arabidopsis thaliana* Tiling Array Express (At-TAX) visualization tools.[9] This enables the community to further investigate the roles of new, uncharacterized transcripts. Raw data were deposited at the NCBI Gene

[9] `http://www.weigelworld.org/resources/microarray/at-tax`

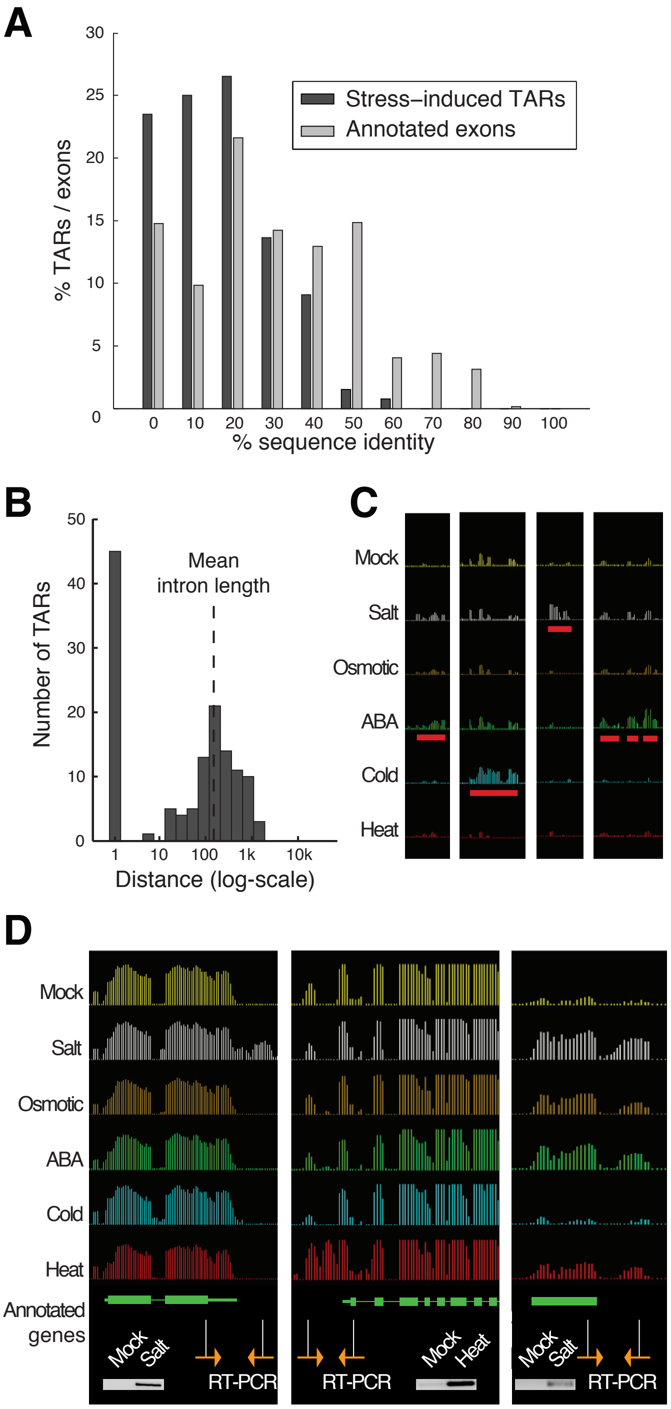


Figure 3.27: Stress-responsive new TARs often overlap with annotated genes. **(A)** Stress-induced new TARs after one hour salt stress were analyzed for their conservation among other plant species. The degree of conservation was assessed based on sequence alignments between *Arabidopsis* and rice, poplar and sorghum (for other stresses see 4.8). **(B)** New TARs identified in salt-stressed plants after one hour were analyzed for their distance to the nearest annotated gene. A distance of 1 indicates overlap of the new TAR with an annotated gene. **(C)** Several examples of new TARs (horizontal red bars) identified under different stress conditions are located in intergenic regions with a distance of more than 500 bp to annotated genes. Vertical bars show the signal of individual tiling array features. **(D)** Examples of stress-induced new TARs that are located close to annotated genes. RT-PCR analysis with one primer located in the annotated gene and one in the new TAR (orange arrows) indicates that these TARs likely represent alternative or new exons of annotated genes rather than independent overlapping transcription units.

Expression Omnibus [45] under the accession code GSE13584.

3.4 Additional Sequence Features for Transcript Mapping (mSTADsp)

In this section we will introduce an extension of mSTAD, dubbed mSTADsp, which incorporates additional DNA sequence features for the improved detection of transcriptionally active regions (TARs). It contains unpublished results obtained with help from Jonas Behr and Gunnar Rätsch.

3.4.1 Accurate Splice Site Recognition from Genomic Sequences

The sequence context around splice sites is probably the most important feature for the *ab initio* prediction of gene structures from genome sequences [140, 155]. Acceptor and donor splice sites delineate the boundaries of internal exons and can be predicted with very high accuracy from the genome sequence alone. Using support vector machines with sophisticated kernels engineered for the classification of genomic sequences [167], Sonnenburg et al. [170] could achieve an accuracy of 99.4% and 99.7% for *Arabidopsis* acceptor and donor splice sites, respectively, as measured by the area under the ROC curve (auROC). Corresponding values for the area under the precision-recall curve (auPRC) were 92.2% and 92.9%, respectively. We therefore incorporated these highly accurate splice site predictions into our segmentation method for tiling array hybridization data with the aim to improve the recognition of TARs in the *Arabidopsis thaliana* genome. Moreover, in contrast to many other types of sequence information useful for gene prediction, splice sites do not only occur in protein-coding genes. Hence, exploiting these sequence features in mSTADsp was not expected to abolish its ability to recognize non-coding transcripts.

For the genome-wide splice site predictions used for the following experiments and kindly provided by Jonas Behr, accuracy was very similar to that reported in Sonnenburg et al. [170] (auROC 99.3% and 99.6% for acceptor and donor splice sites, respectively; auPRC 90.6% and 92.0%, respectively). In addition, Jonas Behr generated a second set of splice site predictions, which were based on a locally restricted sequence context around intron boundaries. While originally around each splice site sequences spanning 81 nt and 60 nt of the adjacent exon and intron, respectively, had been used, the restricted splice site predictors examined only 40 nt of intronic sequence and 10 nt of exonic sequence. These locally restricted splice site predictions were generated with the goal of minimizing biased detection of coding transcripts. However, as a consequence of the smaller sequence windows, accuracy also deteriorated (auROC 98.5% and 99.3%; auPRC 79.4% and 83.4% for acceptor and donor splice sites, respectively).

Before we incorporated splice site predictions as features into our segmentation method, we assessed their predictive power in the context of segmentation with a strand-insensitive model and in comparison to the hybridization signal. For use in strand-insensitive segmentation, the original splice site predictions had to be processed in two ways. First, we mapped them into the probe grid, i.e., obtained only a single value for all splice site predictions between two adjacent tiling probe centers. Essentially, this indicates how likely a splice site is encountered between two consecutive tiling probes. Second, splice site predictions from both strands of the genome had to be reconciled with mSTAD's strand-insensitive segmentation model. We hence combined Watson-strand donor signals and Crick-Strand acceptors to obtain a single (strand-insensitive) score for intron start points; similarly a score for intron end points from Watson-strand acceptors and Crick-strand donors (see also Section 2.7.3 for details). The predictive power was evaluated for each feature individually by means of ROC and precision-recall (PRC) analysis. More specifically, we assessed how informative the hybridization feature was for distinguishing exonic probes from background. For the splice site features, we assessed how well exon-intron and intron-exon transitions, respectively, could be distinguished from transitions between consecutive exon probes, consecutive intron probes and adjacent intergenic probes without annotated splice sites in between (Table 3.9). Lower PRC values for splice site features reflect a more unbalanced prediction problem than, e.g., exon probe recognition. Differences between these accuracy values and those originally obtained for the splice site classifier are likely due to discarding some information about strand and position when mapping splice site predictions into the tiling probe grid.

Feature	auROC	auPRC
Hybridization signal	0.82	0.80
Exon-intron signal		
w50	0.94	0.46
w141	0.97	0.63
Intron-exon signal		
w50	0.93	0.45
w141	0.97	0.63

Table 3.9: Predictive power of sequence and hybridization signals. Splice site features were derived from SVM predictions made with small (w50) and large (w141) sequence windows (see text). auROC denotes the area under the ROC curve, auPRC the area under the precision-recall curve.

3.4.2 Extending the Segmentation Model to Exploit Splice Site Features

As splice sites demarcate intron boundaries, learning from splice site features should be associated with transitions in mSTAD's original state model (Fig. 3.20). To devise a new sequence- and hybridization-based segmentation method, dubbed mSTADsp, we therefore extended the state model by introducing additional splice site states between hybridization states (see Fig. 3.28 for a simplified and Fig. 2.7 for a more comprehensive illustration).

We derived features for mSTADsp in analogy to the extension of the state model discussed above (see Fig. 3.29 for an illustration and Section 2.7.3 for more details). While hybridization states only exploited hybridization signals as a feature for learning, in splice site states this feature was ignored and instead learning was based on splice site prediction scores.

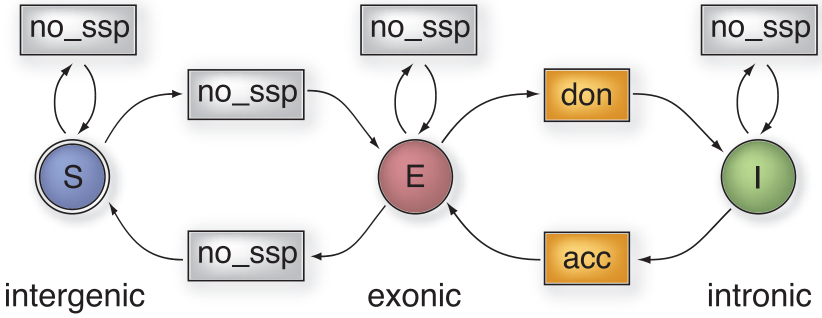


Figure 3.28: Simplified state model employed in mSTADsp. Shown is a submodel with exon and intron states for only one of $Q = 10$ discrete expression levels. States corresponding to hybridization signals are depicted by circles (S, E, I) and splice site prediction states by rectangles (don, acc, no_ssp). It captures the idea that strong splice donor signals are expected between exons and introns of transcripts originating from the Watson strand, likewise strong acceptor signals between intron and exon probes. Ideally, there are no strong splice site signals at the remaining transitions between hybridization states (see also Fig. 2.7 for a more comprehensive state model).

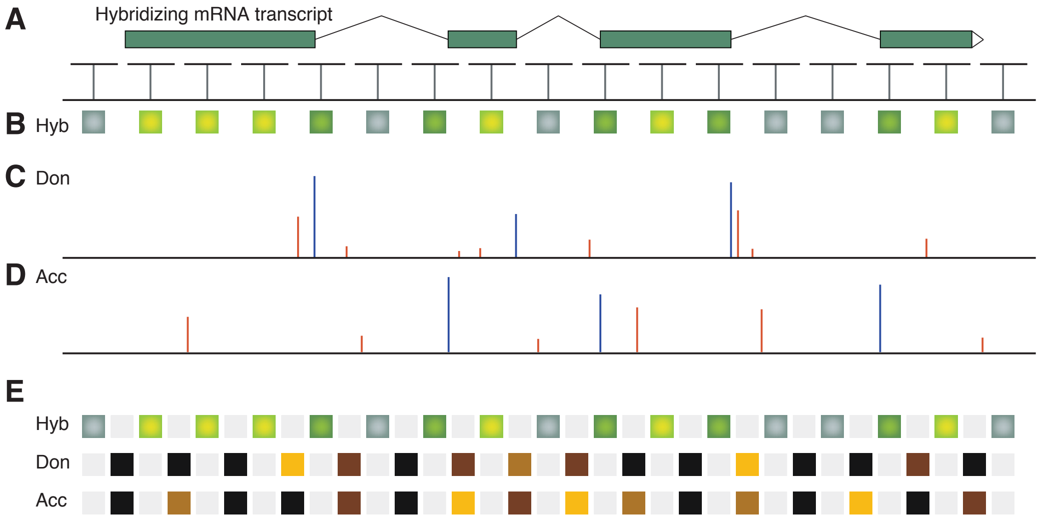


Figure 3.29: Hybridization and sequence-based features utilized by mSTADsp (for simplicity considering only the Watson genome strand). **(A)** Schematic of a spliced mRNA transcript hybridizing to the tiling array, exons drawn as green bars and introns as black lines. **(B)** Hybridization intensity for the probes in the corresponding tiling path, bright green indicating strong signal, dark green intermediate and grey low signal intensity. **(C)** Splice donor signals predicted from the genome sequence. Blue bars correspond to true donor sites, red bars to other candidate sites with the same consensus dinucleotide. Prediction score is indicated by bar height. **(D)** Splice acceptor predictions; colors as for splice donor predictions. **(E)** Schematic view of mSTADsp's feature matrix combining hybridization and splice site signals. For the "Don" and "Acc" features, orange corresponds to high values, brown to intermediate ones and black to low ones. Gray squares indicate that hybridization features were not evaluated in splice site states, whereas splice site features were ignored in hybridization states.

When extending the original model, we had to make the choice whether to keep a strand-insensitive model (motivated from the strand-insensitive nature of our hybridization data), or whether to devise a strand-consistent model. The latter option can be motivated by the fact that splice site predictions are strand-specific. It has the advantage of ensuring that a predicted transcript is consistent in the sense that all its donor and acceptor sites are located on the same genomic strand. However, the number of states in such a model almost doubles compared to the strand-insensitive version. We therefore restricted ourselves to the simpler model, as the number of states proportionally increases computational costs for training and prediction. Nevertheless, additional computational experiments comparing both models would be necessary to exactly quantify the potential accuracy gain associated with the more complex strand-consistent model.

3.4.3 Sequence-Based Splice Site Predictions Improve Exon and Intron Recognition

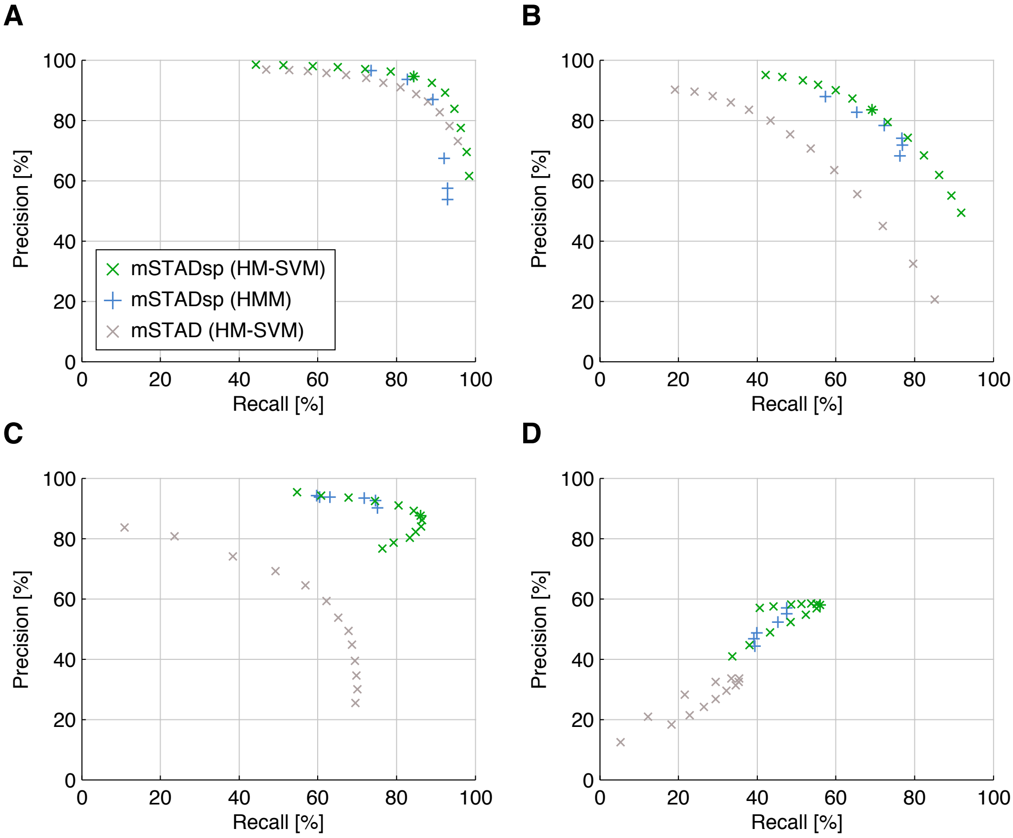


Figure 3.30: Performance improvement of mSTADsp over mSTAD illustrated by precision-recall curves. For mSTADsp, the HM-SVM approach is compared to an equivalent HMM trained generatively. Different trade-offs between precision and recall were obtained similarly as for mSTAD by manipulating the trained segmentation model (see Fig. 3.19, the performance obtained directly after training is indicated by an additional cross). **(A)** Probe-level evaluation. **(B)** Exon-level evaluation **(C)** Intron-level evaluation **(D)** Evaluation of exon-boundary predictions. For definition of precision and recall on the respective evaluation levels see Fig. 3.19.

For assessing the accuracy of TAR predictions made by mSTADsp, we conducted computational experiments very similar to the ones described in Section 3.3.2. We trained mSTADsp discriminatively with the HM-SVM algorithm as well as generatively with HMM training and compared the resulting predictions to those made with mSTAD without splice site features (Fig. 3.30). Incorporating splice site features generated from large sequence windows (w141) resulted in significantly more accurate predictions on probe level, exon level, intron level and exon boundary level. The most considerable improvement was observed for intron recognition, which is not surprising, as all introns are flanked by splice sites, whereas the same is only true for internal, but not terminal exons. Also the correctness of exon boundaries (with respect to tiling probe resolution) increased substantially, now reaching an average of precision and recall of 57% on the test set compared to 34% observed for mSTAD without splice site features — a relative improvement of 67%. We furthermore observed that mSTADsp HM-SVM made consistently better predictions than mSTADsp HMM across all evaluation levels.

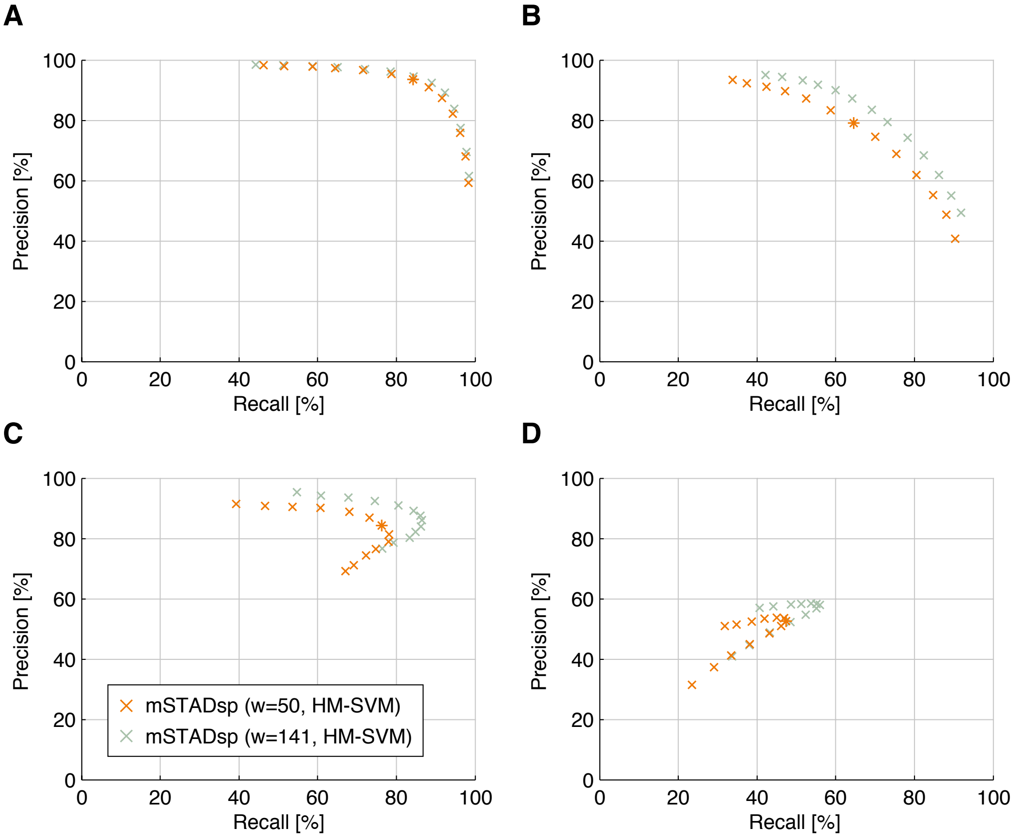


Figure 3.31: Dependency of mSTADsp's accuracy on the size of the local sequence context exploited for splice site prediction. Precision-recall curves show the performance of mSTADsp with splice site features based on a sequences of length 50 nt relative to another mSTAD instance which utilized predictions on 141 nt sequences (Different trade-offs between precision and recall were obtained as before and an additional cross indicates the original performance). **(A)** Probe-level evaluation. **(B)** Exon-level evaluation **(C)** Intron-level evaluation **(D)** Evaluation of exon-boundary predictions. For definition of precision and recall on the respective evaluation levels see Fig. 3.19.

When splice site features were computed from a smaller local sequence context (mSTADsp w50), the resulting TAR predictions were considerably less accurate than those based on splice site prediction with large sequence windows (mSTADsp w141), but nevertheless of substantially higher accuracy than TARs predicted with mSTAD (Fig. 3.31).

A comparative performance assessment of mSTADsp w50, mSTADsp w141 and mSTAD with respect to gene expression levels further supports a clear gain in accuracy of intron and exon (boundary) predictions. Moreover, it revealed that segmentation accuracy improves more strongly for lower gene expression levels, but still helps for the most highly expressed genes (Fig. 3.32 A,B).

Noting that the test set of regions around full-length cDNA-confirmed genes exhibited a biased expression distribution (Fig. 3.32 C), we re-calculated precision and recall aiming to correct for this bias. Precision and recall were calculated separately for each expression level and subsequently averaged across levels. The resulting values are expected to closely reflect prediction accuracy on a genome-wide scale. We observed improvements in recall rates of up to 54%, 66%, and 45% on exon, exon boundary and intron level, respectively, relative to those for TARS predicted by mSTAD without sequence features. The corresponding precision estimates increased by 5%, 109%, and 73%, respectively (Fig. 3.32 A,B). That the increase in precision on exon level appears very small is a consequence of a rather unstable estimate of mSTAD's precision for genes from the lowest expression level, for which its sensitivity is almost 0% (Fig. 3.32 A,B).

3.4.4 TARs Evaluated at Single-Nucleotide Resolution

	predicted intron correct	≥ 1 splice site correct
mSTADsp w50	60.0%	80.6%
mSTADsp w141	68.7%	81.7%

Table 3.10: Intron accuracy of mSTADsp at single-nucleotide resolution. We evaluated 3,778 and 4,336 introns of TARs predicted by mSTADsp w141 and mSTADsp w50, respectively, in genomic regions around 1000 full-length cDNA-confirmed genes (same test set as used before). The first column indicates how many of the predicted introns were annotated with identical start and end point; the second column contains the number of predicted introns for which at least one splice site was also present in an annotated gene.

The fact that splice site predictions have single-nucleotide precision allowed us to resolve TAR boundaries beyond the resolution of the tiling array. Intron boundaries in predicted TARs were mapped back to the position of the splice site between the two adjacent tiling probes for which the highest score was obtained. For the boundaries of terminal exons, we chose the midpoint of the two adjacent tiling probe positions, as splice site information cannot be used to map the ends of TARs. TAR ends are hence not expected to be accurate at the nucleotide level. Intron boundaries, in contrast, could be evaluated at this level and were found to be correct in the majority of predicted cases (Table 3.10). This evaluation

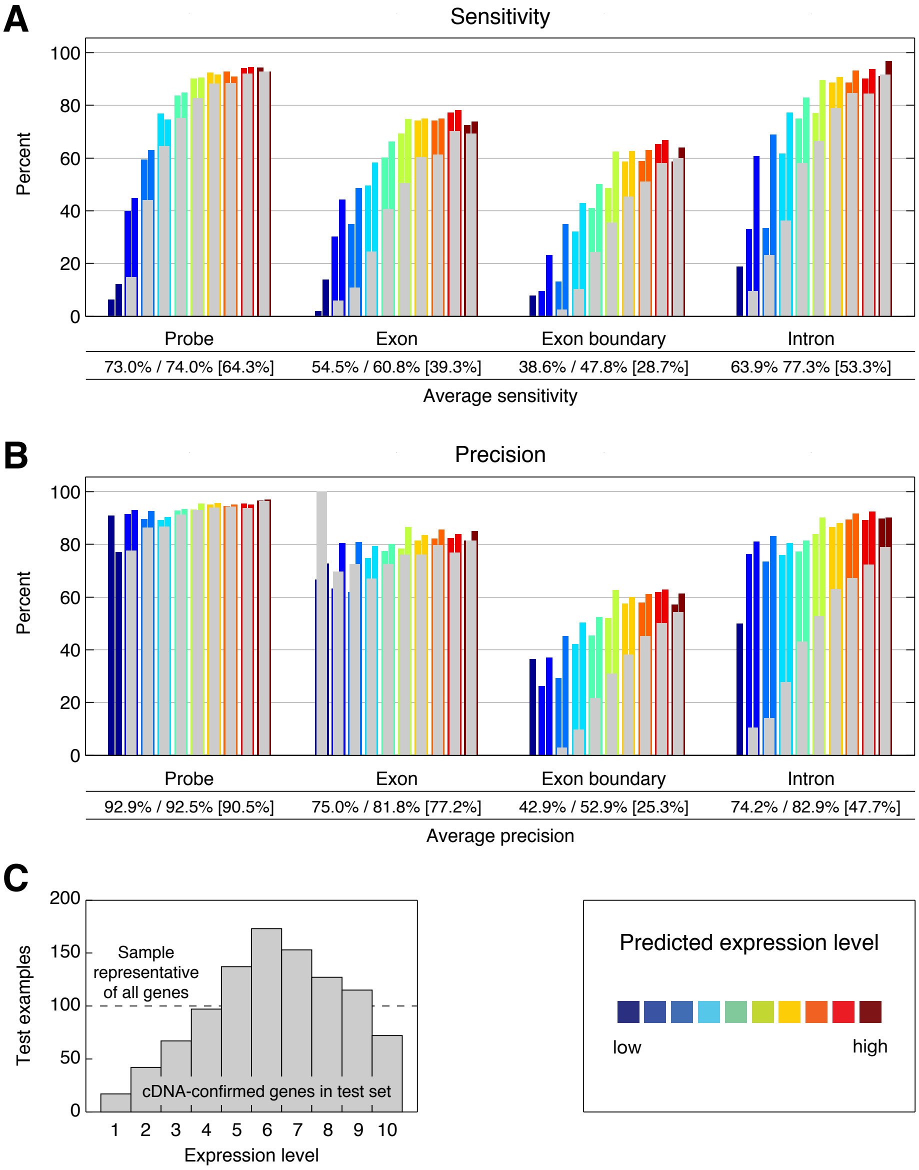


Figure 3.32: Performance of mSTADsp relative to that of mSTAD by gene expression level. Sensitivity (same as recall) **(A)** and precision **(B)** were evaluated for mSTADsp using splice site features computed over small (w50) and large (w141) sequence windows (left and right bars of the same color, respectively). Values for mSTAD without splice site features are indicated by grey bars in front. Average performance is indicated below (mSTADsp w50 / mSTADsp w141 [mSTAD], see main text for details). These values were obtained as averages per expression level to correct for the expression bias of the test set. For definitions of precision and recall (i.e., sensitivity) on probe, exon, exon boundary and intron level see Fig. 3.19. **(C)** Expression bias of full-length cDNA-confirmed genes used for evaluation relative to all annotated genes.

also shows that mSTADsp's intron accuracy strongly depends on the accuracy of the underlying splice site predictions.

To my knowledge, only one other method has been published which combines hybridization and sequence features to predict transcriptionally active regions. ARTADE by Toyoda and Shinozaki [183] uses a Markov chain model to resolve exon-intron structures extending from spots of strong hybridization signals in both, 3' and 5' direction. It models the tiling probe signals together with nucleotide content at single-nucleotide resolution in a small window (10 nt) around splice sites. In contrast to mSTADsp it requires strand-specific hybridization data, which complicates a direct comparison to mSTADsp in terms of prediction accuracy.

When ARTADE was evaluated on a single strand of the genome in regions around ~ 600 full-length cDNAs, it was reported to predict 54.6% of all introns correctly and for 82.2% of the predicted introns, at least one splice site was confirmed by cDNAs [183]. Although these values are not directly comparable to those obtained with mSTADsp (Table 3.10), the fact that a very similar evaluation strategy was used suggests that the accuracy of our method is probably not worse than that of ARTADE. In this context, it is further worth mentioning that strand-specific hybridization data exploited by ARTADE is slightly more informative than the hybridization data used in this work. Additionally, evaluation on a single strand, as done in Toyoda and Shinozaki [183], mitigates potential false-positive errors due to so-called shadow effects caused by transcripts originating from the opposite strand. (Many gene finders account for this effect by employing a shadow model [e.g., 63, 172]). Furthermore, ARTADE's TAR predictions appear to be optimized towards high specificity rather than high sensitivity: genes for which significant expression could not be detected in tiling array data were excluded from ARTADE's test set, and on a whole-genome scale, TARs overlapping by at least 1 nt with known genes, were only predicted for a minority of ~ 27,000 annotated genes [183]. In contrast to that, mSTADsp w141 and mSTADsp w50 predicted TARs that overlap with 90.5% and 95.0% of full-length cDNA-confirmed test genes respectively, and these genes were not pre-filtered by expression support. On a whole-genome scale, ~ 81% and ~ 71% of the 32,805 genes annotated in TAIR8 [178] overlapped with TARs predicted by mSTADsp w141 and mSTADsp w50, respectively. Considering all these arguments, we expect that in this indirect performance comparison ARTADE's performance appears better than it would in a direct comparison on the same test data. Finally, this preliminary result prompts us to speculate that the very high accuracy of the splice site predictions utilized by our method probably has a stronger effect on accuracy than mSTADsp's potential disadvantage of a strand-unspecific model.

3.5 Discussion of Our Normalization and Segmentation Methods

While the previous results sections highlighted the advantages of our newly developed methods, we will discuss some of their limitations in the following and point out possibilities for future extensions.

3.5.1 Limitations and Extensibility of the Proposed Methods

With our newly developed method for the **detection of polymorphic regions** from resequencing array data, mPPR (see Section 3.1), we revealed a large additional fraction of polymorphisms unidentifiable by existing SNP calling techniques (see Section 3.1.2). This is not primarily due to relaxing the stringency of these predictions compared to SNP calling. The resulting PR predictions, however, are less informative than SNP calls since they can essentially contain any type of sequence variant, but we cannot exactly specify which. Furthermore, they only approximately indicate the location of the underlying polymorphisms. Although PR localization is rather precise (Fig. 3.3), it nevertheless remains challenging to assign alleles across accessions. The main reason for this is that prediction boundaries can vary slightly across accessions even if the underlying polymorphisms are the same. Together, these properties severely limit the possibilities for subsequent analyses. For instance, a small PR in the coding region of a gene could either reflect a synonymous SNP without effect on the encoded protein, a nonsense mutation resulting in a truncated protein, or a small deletion causing a shift in the reading frame, etc. The physiological outcome and evolutionary consequence of these sequence variants is extremely different, highlighting the limited use of PR data for further biological analyses.

Even though huge resequencing array data sets exist for a few other organisms, array-based resequencing is already much more expensive than next-generation sequencing (NGS) technologies and the array data has several disadvantages compared to sequence data. This limits the scope of application of mPPR and as NGS-based resequencing projects progress, the PR data generated thus far will soon lose its relevance. Nevertheless the algorithm might — with modifications — be applicable to the analysis of other types of data, such as array-based comparative genome hybridization (aCGH) [e.g., 134, 164]. Currently, aCGH is still a relatively cost-effective technique, especially for clinical applications, and a very active field of research. The detection of deletions and copy number variants from NGS resequencing data might be another route for developing mPPR further.

Our **transcript normalization** method (see Section 3.2) effectively removes probe sequence bias (Figs. 3.13 and 3.14) and thereby improves the difference between the exon and the background intensity distributions (see Section 3.2.5). Nevertheless, it has several limitations, some of which could possibly be overcome by future extensions. First of all, the transcript normalization model assumes that the ideal transcript intensity is constant. However, we know that this is not the case for real data due to an amplification bias (Fig. 3.15). We even demonstrated that normalization accuracy can be improved when using a more flexible transcript intensity model (see Section 3.2.4). However, this model additionally requires information on transcript structures and can thus not be incorporated in a preprocessing routine that is independent of existing gene annotations or subsequent transcript identification methods.

Another shortcoming of the transcript normalization method is its inability to deal with the additional information contained in mismatch (MM) probes in a satisfactory manner. In the current Affymetrix Tiling Array 1.0R array for *Arabidopsis*, half the array is occupied by these mismatch probes and in previous work, they have been successfully used to estimate (probe sequence-dependent) unspecific background hybridization and to improve exon-background separation [e.g., 120, 195]. Including MM intensities directly as additional features into our transcript normalization procedure did, however, not improve the results. Future experiments are necessary to explore, whether there are alternatives on how to incorporate these data such that they improve normalization accuracy.

Finally, although we demonstrated the benefits of modeling different intensity quantiles individually (see Section 3.2.5), this complete separation seems artificial and may even be suboptimal in terms of normalization accuracy. For instance, we cannot guarantee that the joint normalization is monotonic across intensity quantiles for a fixed probe sequence. Transferring knowledge across submodels could be realized via a joint regularization term that couples the individual solutions. Whether solving the larger joint problem justifies possible accuracy gains remains to be investigated.

Currently, the most obvious disadvantage of our **transcript identification method**, mSTAD (see Section 3.3), is that it is designed to segment data from each sample separately. Due to the limited accuracy of predicted segment boundaries, it is difficult to compare predictions across samples to distinguish biologically relevant differences from ones which result from noise in the data and from the stochastic nature of the algorithm (e.g., Fig. 3.25). Ideally, several samples should be segmented in conjunction to reduce the uncertainty at segment boundaries; the remaining differences should then mainly reflect differential expression of genes or transcript isoforms. This is, however, not easily formalized as a label sequence learning problem (with a small set of labels) and hence poses a theoretical challenge. A practically more viable alternative might be an additional second layer of segmentation models that exploit as additional features the output of the first segmentation layer obtained for other samples (e.g., similar to the two-layered SNP calling method proposed in Clark et al. [29]).

A more general limitation results from the array design itself. The probe density along the chromosome imposes a limit on segmentation accuracy; in case of the *Arabidopsis* tiling array analyzed here, many exons and especially introns are covered by very few (0 – 2) tiling probes and are thus difficult to detect accurately.

Finally, mSTAD's state model (Fig. 3.20), which treats genes separately depending on their expression level, contributes substantially to high segmentation accuracy (Fig. 3.21). However, the discrete model of differential expression yields only rough expression estimates, which are of limited use for identifying differentially expressed genes or exons. We therefore had to separately derive accurate gene expression measurements and apply a separate statistical test to identify differentially expressed genes and TARs, which appears suboptimal, e.g., in comparison to Huber et al. [72].

In case of **transcript identification using sequence features** with mSTADsp (see Section 3.4), many of mSTAD's limitations apply as well, with the exception that sequence features, such as splice sites, allow resolving gene structures beyond the resolution of the tiling array (see Section 3.4.4). Further possibilities to extend mSTADsp appear promising: i) It is relatively straight-forward to implement a strand-consistent state model (as discussed in Section 3.4.2). ii) More sequence-based features could be employed; for instance, transcription start site predictions [169] could help improve the accurate detection of gene boundaries — currently a weakness of mSTADsp. iii) Taking a semi-Markov approach would allow us to additionally model segment properties, such as length distributions or nucleotide content to further improve prediction accuracy [137]. This has already proven useful in the context of sequence-based gene finding [e.g., 155, 172].

General limitations of the methods presented in this dissertation are first of all related to their complexity. Training an HM-SVM involves solving a linear or quadratic optimization problem, which typically comprises thousands of variables (see Sections 2.4 and 2.4.8). For this, we used commercial optimization software, which facilitates training on many more examples than would be possible with freely available software. Together with the fact that the proposed methods are implemented in Matlab, this impedes their wider application. Furthermore, to achieve good generalization performance it is necessary to tune several hyperparameters of the learning algorithms. Doing this in a systematic model selection is very demanding in terms of computation time. To address these issues, we intend to provide a version of the source code that is fully compatible with Octave and facilitates the use of free optimization software. Additionally, we will provide a web server version, which offers preselected hyperparameters and allows external users to profit from our software resources.

Another general shortcoming of our methods relates to the underlying supervised learning algorithms. While they generalize well on unseen data that is represented well in their training set (i.e., if the distributions of training and target data are identical), that does not have to be the case if this assumption is invalid. Usually, the learnt models are too complex to easily understand or predict their behavior on examples that are underrepresented among the available labeled data. For instance, PR predictions were made for the whole genome including repetitive regions and large deletions, which were absent from our training and test data. We therefore had to perform additional evaluations based on further data sources to ensure that predictions were reasonable in these regions as well (see Section 3.1.3).

The modular nature of our methods for transcriptome tiling array analysis was motivated by practical reasons. However, a method that combines transcript normalization and transcript identification may yield better results than obtained by sequential application of normalization (TN) and segmentation (mSTAD). Such a combined approach could potentially model the intensity bias along the transcript and thereby overcome the limitation of constant transcript intensities inherent in TN and mSTAD.

Eventually, although the methods presented in this thesis were originally developed for and tested on *Arabidopsis* tiling arrays, they are, however, not limited in application to these data. Resequencing array data sets similar to the one generated for *Arabidopsis thaliana* also exist for human, mouse, rice and yeast [54, 68, 111, 148], and the prediction of polymorphic regions has already been successfully applied to rice resequencing data [10]. Comprehensive corpora of transcriptome tiling array data have been generated for many model organisms including human, fly, worm, rice, and yeast [28, 35, 86, 102, 108, and others]. Further experiments with these data would be highly desirable to enhance the visibility of our methods and to confirm that their accuracy is not due to peculiarities of a single organism, but rather due to the power of the underlying machine learning methods. As a first step towards this goal, we intend to analyze data for several *C. elegans* tissues and specific cell types collected within the scope of the modENCODE project.[10]

3.5.2 Applicability of the Proposed Methods to RNA-seq Data

Although whole-genome tiling arrays became available for a large research community only recently [117], it seems to be a commonplace among computational biologists that this technology is soon going to be replaced by next-generation sequencing [161]. Ultra-high throughput sequencing of the transcriptome (RNA-seq for short) has already been demonstrated to yield quantitative and structural information at unprecedented resolution [e.g., 118]. Consequently, developing tools for the analysis of RNA-seq data is currently a top priority in bioinformatics. In the following paragraph, I will sketch how our methods for analyzing tiling array data might also be applicable to RNA-seq data.

Despite the fact that tiling arrays and RNA-seq are technologically very dissimilar, some of the problems encountered in the data analysis phase are related: Measurements of transcript quantities with both, RNA-seq and tiling arrays, are confounded by sequence effects. We and others characterized sequence effects of tiling array probes in detail (in Section 3.2, [e.g., 145]). For next-generation sequencing of genomes and transcriptomes, it has also been shown that the efficiency of generating a read depends on the sequence of the target (template) DNA [e.g., 105, 130]. Here, we analyzed the dependency of RNA-seq read coverage within annotated exons on the local GC content in a 25-bp context (Fig. 3.33, see also Fig. 3.13).[11] For this, we only considered positions covered by more than one read. Estimates for GC-poor regions ($< 20\%$) appear very uncertain, possibly due to mismapped reads originating from polyA tails. Elsewhere, we observe an approximately five-fold increase in read coverage with GC content (between 25% and 60%; Fig. 3.33). This finding suggests that accurate transcript quantification may benefit from sequence normalization techniques similar to our transcript normalization method (see Section 3.2).

The identification of new transcripts from RNA-seq data is arguably easier than from tiling array data. When read coverage is sufficient, RNA-seq data will not only facilitate exon-background separation, but spliced reads will also accurately delineate exon-intron boundaries at 1-bp resolution, and even provide direct support for introns [12]. Nevertheless, the structures of transcripts with low expression, which are only sparsely covered by

[10] `http://www.modencode.org`

[11] For generating and mapping the underlying *Arabidopsis* RNA-seq data, I would like to gratefully acknowledge Jun Cao, Stephan Ossowski, Korbinian Schneeberger, Fabio De Bona and Regina Bohnert.

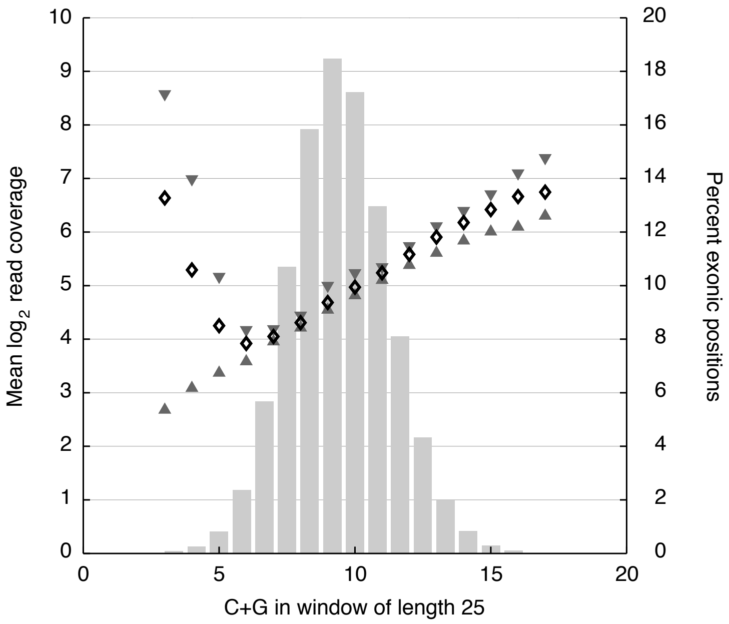


Figure 3.33: Coverage of exons with RNA-seq reads is biased by GC content. Average position-wise log_2-transformed coverage is shown as a function of GC count in a 25 bp window centered on positions that were contained within an exon annotated in TAIR8 [178] and covered by more than one read (black diamonds). Additionally, the minimum and maximum values from an analysis of individual chromosomes are indicated (by gray triangles pointing upward and downward, respectively). Gray bars depict a GC histogram among the positions considered.

RNA-seq reads may be less obvious from the primary data. Importantly, our tiling array-based transcriptome studies indicate, however, that new genes are expected to exhibit relatively low overall expression levels with patterns that are probably restricted to specific conditions or tissues [97]. Only such genes are likely to have escaped previous annotation efforts in extensively studied model organisms such as *Arabidopsis thaliana.* Techniques like our method for tiling array segmentation (mSTAD, see Section 3.3), may hence be well-suited to more accurately resolve the structures of such transcripts. More precisely, we envision several possibilities to exploit RNA-seq data for transcript identification. A straight-forward strategy would be to convert mapped RNA-seq reads into (exon) coverage counts and simply treat them as a 1-bp tiling for the purpose of segmentation. Here, unmappable and repetitive regions could be handled similarly as cross-hybridizing probes. Several extensions of this naive approach are conceivable: First, it appears promising to combine read-coverage features with genomic sequence features as done with mSTADsp (see Section 3.4). Second, it would also be possible to convert spliced reads into an intron coverage feature to further improve intron recognition accuracy. Such approaches require only minor adjustments of mSTAD or mSTADsp. There are however additional properties of RNA-seq data which one might like to exploit, e.g., the connectivity structure imposed by intron reads spanning two splice sites or mate pair information connecting two reads. Devising a segmentation which takes these properties into account is, however, much more difficult. These long-range dependencies are not easily reconciled with the Markov assumption, that the optimal segmentation at a given position only depends on directly adjacent positions. Even when using semi-Markov models, as e.g., proposed in Rätsch and Sonnenburg [137], intron connectivity could be modeled in simple cases, but mate-pair information still poses a challenge. When analyzing RNA-seq data with our segmentation methods, their inability to predict alternative transcripts is a severe limitation as the data is in principle well-suited for the discovery of alternative isoforms. Transcript mapping from RNA-seq data can also be approached from a different perspective, i.e., by using sequence-based gene finding systems which incorporate RNA-seq reads as additional features similar to the way ESTs are typically handled [e.g., 155]. While segmentation approaches like mSTAD and mSTADsp may generally be less accurate than sequence-based gene finders, they might, however, complement such approaches. As mSTAD and mSTADsp make no or very little assumptions about genic DNA sequences, respectively, they are expected to readily detect expressed non-coding transcripts, whereas the models employed in sequence-based gene finders are typically restricted to protein-coding genes.

In conclusion, notwithstanding that we might soon witness “the end for microarrays” [161], the analysis methods presented in this work are rather broadly applicable. With a few modifications, they will be applicable to RNA-seq data. As there exists a pressing need for computational tools to cope with the rapidly increasing output of sequencing instruments, this seems to be a very promising direction for future development.

3.6 Chapter Summary

In this chapter I presented the application of newly developed machine learning-based methods to the analysis of whole-genome microarrays.

More specifically, the application of mPPR (see Section 3.1) to resequencing array data from *Arabidopsis* revealed 240,538 to 361,184 polymorphic regions (PR) per accession (5.3-8.5% of the genome). Although the exact types of contained polymorphisms cannot be specified based on this data, I carefully verified that their location coincided very precisely with known polymorphisms. Additionally, we could even accurately predict large deletions and monomorphic regions that were absent from the training set (Section 3.1.3). When we analyzed the PR distribution around annotated genes, we found striking nonrandom patterns with exons, core promoters and miRNAs exhibiting significantly reduced polymorphism levels (Sections 3.1.4 and 3.1.5). We also found large variation in PR content of genes in different families and that resistance genes (NB-LRR) were exceptional in that they had the highest overall PR level and the largest variation between individual family members (Section 3.1.5). Taken together, these results suggest large differences between the functional gene complement of individual plants

Furthermore, I developed a normalization pipeline (Section 3.2) with a novel transcript normalization method as its core component. Although many methods have been proposed to alleviate probe sequence biases [e.g., 120, 122, 145], comparisons of our transcript normalization to these methods showed that ours is the only one which effectively reduces these biases *and* improves signal-to-noise ratio (as measured by exon-probe recognition on normalized data). Additional experiments indicated that other methods failed in this aspect, probably because they primarily model sequence effects for unspecific binding and only poorly account for specific binding to exons of expressed transcripts. Combining this normalization with a newly developed *de novo* transcript mapping method (Section 3.3) allowed us to recognize expressed transcripts with unprecedented accuracy corroborated by a comparative evaluation including other methods (Section 3.3.2) and by experimental validation (Section 3.3.3). As a result of the application of these methods to a large body of transcriptome tiling array data for *Arabidopsis*, I discovered more than 1,000 new transcriptionally active regions (TARs), per sample. These were absent from the current gene annotation, despite the fact that it is based on previous tiling array experiments. Interestingly, the expression of hundreds of these TARs was found to be significantly altered upon treatment with abiotic stresses, suggesting that potentially genes with a function in adaptation to adverse environments may be included. Extending the transcript mapping method such that it exploits additional sequence features, I demonstrated that the accuracy can be improved even further (Section 3.4). Finally, the applicability of similar methodology to the task of recognizing transcripts from RNA-seq data is discussed as a promising route for future development (Section 3.5.2).

4 Conclusion

In this dissertation, I present state-of-the-art machine learning methods for the analysis of data from whole-genome tiling arrays. These methods were developed specifically for two major applications of tiling arrays, namely transcriptome profiling and polymorphism discovery.

I faced a number of technical challenges associated with the primary microarray data [see also 144]. First, to minimize artifacts due to *cross-hybridization* in regions of repetitive sequence, I created a custom repeat-annotation (Section 2.3.4). From the genome sequence, I identified those repeat classes with the greatest potential of confounding the tiling probe signals. A second issue, especially with transcriptome tiling array data, is a well-known *probe-sequence bias* (Section 3.2): the fact that hybridization properties diverge strongly between different probe sequences has a pronounced effect on the signal read-out. Because reference data was collected within the scope of the resequencing project, I could use it as an indirect means of normalization (Section 3.1.1). In the case of tiling array-based transcriptomics, our newly developed transcript normalization method effectively alleviates the probe sequence bias (Sections 2.3.3 and 3.2) [204]. Additionally, it reduces the variance in the hybridization signal in a way that benefits subsequent transcript identification. This is in sharp contrast to other methods developed for this purpose. The ones I investigated in this work reduce probe sequence effects at the cost of diminishing true transcript signals, which essentially renders them useless as preprocessing routines for transcript identification (Section 3.2.5). Third, we developed methods for the *detection of polymorphic regions* (Sections 2.5 and 3.1) [203], as well as for *transcript identification* (Sections 2.6 and 3.3) [97, 204] from tiling array data, which are able to accurately segment hybridization measurements in the presence of high noise levels that are typical of DNA microarray data. For both applications, accuracy is of prime importance because biological interpretation demands very low false discovery rates.

The methods which we developed to solve these problems are based on state-of-the-art supervised machine learning techniques, including support vector machines (SVMs) [e.g., 6, 153] and hidden Markov SVMs (HM-SVMs) (Section 2.4) [3, 184]. Where competing computational tools were available for comparison, as for transcript identification, I could demonstrate the superior *accuracy* of our methodology (Section 3.3.2). Its predictive power, however, comes at the cost of several limitations (discussed in Section 3.5.1): The *complexity* of our methods, e.g., computationally demanding training procedures, may limit their usability by other researchers. Simplified implementations and a web-server interface will help to alleviate this problem. Thus far, I have tested our methods only on data from a single organism, *Arabidopsis thaliana.* Further experiments using existing data sets from several other organisms are necessary to underline their *general applicability to a diverse set of organisms* and to confirm that their accuracy is not due to the peculiarities of a particular one. The rapid development of next-generation sequencing technology might soon bring “the end for microarrays” [161]. It is therefore of central importance to extend the methodology to *handle sequencing-based polymorphism and transcriptome data* (Section 3.5.2). Fortunately, adjusting them to the particulars of new experimental technologies is greatly facilitated by the fact that they are based on adaptive machine learning techniques. I thus believe that the contributions of this thesis to methodology will outlive the current, sequencing technology-driven revolution in genomics and transcriptomics.

I applied our methods to data from *Arabidopsis* whole-genome arrays to gain insights into the patterns of natural sequence variation (Section 3.1), as well as to reveal transcriptional activity and dynamics (Section 3.3). Both studies were global in the sense that we could exploit, for the first time, data from the *whole genome* of *Arabidopsis thaliana*. *Polymorphic region* (PR) predictions indicated the approximate locations (with a precision of 97%) of a substantial proportion of sequence variants undetected by previous efforts [29, 203]. An estimated 42% of total single-nucleotide polymorphisms (SNPs) were included within PR boundaries in addition to those discovered previously with different methods at a precision of 98% (Section 3.1.2) [29, 203]. Moreover, PRs contained insertion sites and deletions (indels) as well as highly divergent sequence tracts varying in length from a few nucleotides to several thousands (Section 3.1.3). The resulting *Arabidopsis* polymorphism resource was thus the first one to contain small to large indels on a whole-genome scale. This data allowed us to identify genes which are probably nonfunctional in several of the accessions studied, due to SNPs disrupting splicing or truncating the encoded protein, or due to PRs indicating partial or complete gene deletions. We also found surprisingly large differences between the functional gene complement of individual plants (Section 3.1.5) [29, 203]. Our *tiling array-based transcriptome atlas*, At-TAX [97], provided, for the first time, whole-transcriptome measurements for several *Arabidopsis* tissues, developmental stages and stress conditions. It comprises expression profiles for nearly 10,000 genes that were neglected by previous studies because they could not be analyzed with the most widely used gene-centric microarray platform for *Arabidopsis* [97]. Additionally, the results of our *de novo* transcript identification method revealed expressed regions on a genome-wide scale irrespective of gene annotations (Section 3.3.3). A global comparison of polyadenylated and nonpolyadenylated transcripts based on these data suggests that nonpolyadenylated RNAs only make a minor contribution to the *Arabidopsis* transcriptome. This finding is in surprising contrast to studies in human and *Caenorhabditis elegans* where about half of all transcripts are not polyadenylated (Section 3.3.3). Monitoring the *Arabidopsis* transcriptome under abiotic stress conditions, I found several hundred regions that are not annotated as genes but exhibit interesting stress-responsive expression patterns. These not only include candidates for entirely new genes but also previously overlooked exons and alternative transcripts of known genes (Section 3.3.4) [205]. Taken together, our computational tools provide an excellent foundation for further characterizing the *Arabidopsis* transcriptome, in particular the global effects of genetic defects. For instance, investigating mutants impaired in the biogenesis of small RNAs (*se, hyl1, dcl1*), we recently discovered defects in pre-mRNA splicing, microRNA processing and suppression of transcripts originating from transposable elements [96, 98]. In the near future, I expect to see many similar studies creating a demand for our computational tools. In favor of this prospect, Brady and Provart [16] lately highlighted the value of both, the *Arabidopsis* polymorphism inventory and the At-TAX atlas, as publicly available resources for generating hypotheses in plant biology.

Bibliography

[1] M. D. Adams, J. M. Kelley, J. D. Gocayne, M. Dubnick, M. H. Polymeropoulos, H. Xiao, C. R. Merril, A. Wu, B. Olde, and R. F. Moreno. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, 252(5013):1651–6, 1991.

[2] J. M. Alonso and J. R. Ecker. Moving forward in reverse: Genetic technologies to enable genome-wide phenomic screens in Arabidopsis. *Nature Reviews Genetics*, 7(7):524–36, 2006.

[3] Y. Altun, I. Tsochantaridis, and T. Hofmann. Hidden Markov support vector machines. *Proceedings of the ICML*, 2003.

[4] E. G. Bakker, C. Toomajian, M. Kreitman, and J. Bergelson. A genome-wide survey of R gene polymorphisms in Arabidopsis. *The Plant Cell*, 18(8):1803–18, 2006.

[5] J. A. Bedell, M. A. Budiman, A. Nunberg, R. W. Citek, D. Robbins, J. Jones, E. Flick, T. Rholfing, J. Fries, et al. Sorghum genome sequencing by methylation filtration. *PLoS Biology*, 3(1):e13, 2005.

[6] A. Ben-Hur, C. S. Ong, S. Sonnenburg, B. Schölkopf, and G. Rätsch. Support vector machines and kernels for computational biology. *PLoS Computational Biology*, 4(10):e1000173, 2008.

[7] A. Bernal, K. Crammer, A. Hatzigeorgiou, and F. Pereira. Global discriminative learning for higher-accuracy computational gene prediction. *PLoS Computational Biology*, 3(3):e54, 2007.

[8] P. Bertone, V. Stolc, T. E. Royce, J. S. Rozowsky, A. E. Urban, X. Zhu, J. L. Rinn, W. Tongprasit, M. Samanta, et al. Global identification of human transcribed sequences with genome tiling arrays. *Science*, 306(5705):2242–6, 2004.

[9] K. Birnbaum, D. E. Shasha, J. Y. Wang, J. W. Jung, G. M. Lambert, D. W. Galbraith, and P. N. Benfey. A gene expression map of the Arabidopsis root. *Science*, 302(5652):1956–60, 2003.

[10] R. Bohnert, K. Childs, G. Zeller, R. M. Clark, C. R. Buell, D. Weigel, and G. Rätsch. Genome-wide detection of polymorphic regions in domesticated rice. *In preparation*, 2009.

[11] B. M. Bolstad, R. A. Irizarry, M. Astrand, and T. P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–93, 2003.

[12] F. De Bona, S. Ossowski, K. Schneeberger, and G. Rätsch. Optimal spliced alignments of short sequence reads. *Bioinformatics*, 24(16):i174–80, 2008.

[13] G. M. Borchert, W. Lanier, and B. L. Davidson. RNA polymerase III transcribes human microRNAs. *Nature Structural & Molecular Biology*, 13(12):1097–101, 2006.

[14] J. O. Borevitz, D. Liang, D. Plouffe, H.-S. Chang, T. Zhu, D. Weigel, C. C. Berry, E. Winzeler, and J. Chory. Large-scale identification of single-feature polymorphisms in complex genomes. *Genome Research*, 13(3):513–23, 2003.

[15] J. O. Borevitz, S. P. Hazen, T. P. Michael, G. P. Morris, I. R. Baxter, T. T. Hu, H. Chen, J. D. Werner, M. Nordborg, et al. Genome-wide patterns of single-feature polymorphism in Arabidopsis thaliana. *Proceedings of the National Academy of Sciences of the USA*, 104(29):12057–62, 2007.

[16] S. Brady and N. Provart. Web-queryable large-scale data sets for hypothesis generation in plant biology. *The Plant Cell*, 2009.

[17] S. Brenner, M. Johnson, J. Bridgham, G. Golda, D. H. Lloyd, D. Johnson, S. Luo, S. McCurdy, M. Foy, et al. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nature Biotechnology*, 18(6):630–4, 2000.

[18] M. P. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Ares, and D. Haussler. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences of the USA*, 97(1):262–7, 2000.

[19] M. Brudno, S. Malde, A. Poliakov, C. B. Do, O. Couronne, I. Dubchak, and S. Batzoglou. Glocal alignment: Finding rearrangements during alignment. *Bioinformatics*, 19 Suppl 1:i54–62, 2003.

[20] C. Burge and S. Karlin. Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology*, 268(1):78–94, 1997.

[21] W. Busch and J. U. Lohmann. Profiling a plant: Expression analysis in Arabidopsis. *Current Opinion in Plant Biology*, 10(2):136–41, 2007.

[22] M. E. Chaboute, N. Chaubet, B. Clement, C. Gigot, and G. Philipps. Polyadenylation of histone H3 and H4 mRNAs in dicotyledonous plants. *Gene*, 71(1):217–23, 1988.

[23] M. E. Chaboute, N. Chaubet, C. Gigot, and G. Philipps. Histones and histone genes in higher plants: Structure and genomic organization. *Biochimie*, 75(7):523–31, 1993.

[24] N. Chaubet, M. E. Chaboute, B. Clément, M. Ehling, G. Philipps, and C. Gigot. The histone H3 and H4 mRNAs are polyadenylated in maize. *Nucleic Acids Research*, 16(4):1295–304, 1988.

[25] M. Chee, R. Yang, E. Hubbell, A. Berno, X. C. Huang, D. Stern, J. Winkler, D. J. Lockhart, M. S. Morris, and S. P. Fodor. Accessing genetic information with high-density DNA arrays. *Science*, 274 (5287):610–4, 1996.

[26] J. A. Chekanova, B. D. Gregory, S. V. Reverdatto, H. Chen, R. Kumar, T. Hooker, J. Yazaki, P. Li, N. Skiba, et al. Genome-wide high-resolution mapping of exosome substrates reveals hidden features in the Arabidopsis transcriptome. *Cell*, 131(7):1340–53, 2007.

[27] W. Chen, N. J. Provart, J. Glazebrook, F. Katagiri, H.-S. Chang, T. Eulgem, F. Mauch, S. Luan, G. Zou, et al. Expression profile matrix of Arabidopsis transcription factor genes suggests their putative functions in response to environmental stresses. *The Plant Cell*, 14(3):559–74, 2002.

[28] J. Cheng, P. Kapranov, J. Drenkow, S. Dike, S. Brubaker, S. Patel, J. Long, D. Stern, H. Tammana, et al. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science*, 308(5725): 1149–54, 2005.

[29] R. M. Clark, G. Schweikert, C. Toomajian, S. Ossowski, G. Zeller, P. Shinn, N. Warthmann, T. T. Hu, G. Fu, et al. Common sequence polymorphisms shaping genetic diversity in Arabidopsis thaliana. *Science*, 317(5836):338–42, 2007.

[30] Encode Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146):799–816, 2007.

[31] International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437(7063): 1299–320, 2005.

[32] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 1995.

[33] O. Couronne, A. Poliakov, N. Bray, T. Ishkhanov, D. Ryaboy, E. Rubin, L. Pachter, and I. Dubchak. Strategies and tools for whole-genome alignments. *Genome Research*, 13(1):73–80, 2003.

[34] D. J. Cutler, M. E. Zwick, M. M. Carrasquillo, C. T. Yohn, K. P. Tobin, C. Kashuk, D. J. Mathews, N. A. Shah, E. E. Eichler, et al. High-throughput variation detection and genotyping using microarrays. *Genome Research*, 11(11):1913–25, 2001.

[35] L. David, W. Huber, M. Granovskaia, J. Toedling, C. J. Palm, L. Bofkin, T. Jones, R. W. Davis, and L. M. Steinmetz. A high-resolution map of transcription in the yeast genome. *Proceedings of the National Academy of Sciences of the USA*, 103(14):5320–5, 2006.

[36] E. Dawson, Y. Chen, S. Hunt, L. J. Smink, A. Hunt, K. Rice, S. Livingston, S. Bumpstead, R. Bruskiewich, et al. A SNP resource for human chromosome 22: Extracting dense clusters of SNPs from the genomic sequence. *Genome Research*, 11(1):170–8, 2001.

[37] D. DeCaprio, J. P. Vinson, M. D. Pearson, P. Montgomery, M. Doherty, and J. E. Galagan. Conrad: Gene prediction using conditional random fields. *Genome Research*, 17(9):1389–98, 2007.

[38] K. Denby and C. Gehring. Engineering drought and salinity tolerance in plants: Lessons from genome-wide expression profiling in Arabidopsis. *Trends in Biotechnology*, 23(11):547–52, 2005.

[39] J. R. Dinneny, T. A. Long, J. Y. Wang, J. W. Jung, D. Mace, S. Pointer, C. Barron, S. M. Brady, J. Schiefelbein, and P. N. Benfey. Cell identity mediates the response of Arabidopsis roots to abiotic stress. *Science*, 320(5878):942–5, 2008.

[40] C. B. Do, D. A. Woods, and S. Batzoglou. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*, 22(14):e90–8, 2006.

[41] Z. Dominski and W. F. Marzluff. Formation of the 3' end of histone mRNA: Getting closer to the end. *Gene*, 396(2):373–90, 2007.

[42] J. Du, J. S. Rozowsky, J. O. Korbel, Z. D. Zhang, T. E. Royce, M. H. Schultz, M. Snyder, and M. Gerstein. A supervised hidden Markov model framework for efficiently segmenting tiling array data in transcriptional and chIP-chip experiments: systematically incorporating validated biological knowledge. *Bioinformatics*, 22(24):3016–24, 2006.

[43] B. P. Durbin, J. S. Hardin, D. M. Hawkins, and D. M. Rocke. A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics*, 18 Suppl 1:S105–10, 2002.

[44] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis: Probabilistic models of protein and nucleic acids.* Cambridge University Press, 7th edition, 1998.

[45] R. Edgar, M. Domrachev, and A. E. Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–10, 2002.

[46] R. C. Edgar. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–7, 2004.

[47] J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, et al. Real-time DNA sequencing from single polymerase molecules. *Science*, 323(5910):133–8, 2009.

[48] N. Fahlgren, M. D. Howell, K. D. Kasschau, E. J. Chapman, C. M. Sullivan, J. S. Cumbie, S. A. Givan, T. F. Law, S. R. Grant, et al. High-throughput sequencing of Arabidopsis microRNAs: Evidence for frequent birth and death of MIRNA genes. *PLoS ONE*, 2(2):e219, 2007.

[49] N. V. Fedoroff. Cross-talk in abscisic acid signaling. *Science Signaling*, 2002(140):RE10, 2002.

[50] S. Fowler and M. F. Thomashow. Arabidopsis transcriptome profiling indicates that multiple regulatory pathways are activated during cold acclimation in addition to the CBF cold response pathway. *The Plant Cell*, 14(8):1675–90, 2002.

[51] J. M. Franco-Zorrilla, A. Valli, M. Todesco, I. Mateos, M. I. Puga, I. Rubio-Somoza, A. Leyva, D. Weigel, J. A. García, and J. Paz-Ares. Target mimicry provides a new mechanism for regulation of microRNA activity. *Nature Genetics*, 39(8):1033–7, 2007.

[52] K. A. Frazer, L. Pachter, A. Poliakov, E. M. Rubin, and I. Dubchak. VISTA: Computational tools for comparative genomics. *Nucleic Acids Research*, 32(Web Server issue):W273–9, 2004.

[53] K. A. Frazer, C. M. Wade, D. A. Hinds, N. Patil, D. R. Cox, and M. J. Daly. Segmental phylogenetic relationships of inbred mouse strains revealed by fine-scale analysis of sequence variation across 4.6 Mb of mouse genome. *Genome Research*, 14(8):1493–500, 2004.

[54] K. A. Frazer, E. Eskin, H. M. Kang, M. A. Bogue, D. A. Hinds, E. J. Beilharz, R. V. Gupta, J. Montgomery, M. M. Morenzoni, et al. A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. *Nature*, 448(7157):1050–3, 2007.

[55] B. J. Frey, Q. D. Morris, W. Zhang, N. Mohammad, and T. R. Hughes. GenRate: a generative model that finds and scores new genes and exons in genomic microarray data. *Pacific Symposium on Biocomputing*, pages 495–506, 2005.

[56] B. J. Frey, Q. D. Morris, and T. R. Hughes. GenRate: a generative model that reveals novel transcripts in genome-tiling microarray data. *Journal of Computational Biology*, 13(2):200–14, 2006.

[57] R. Giegerich, C. Meyer, and P. Steffen. A discipline of dynamic programming over sequence data. *Science of Computer Programming*, 2004.

[58] D. Gilbert and A. Rechtsteiner. Comments on sequence normalization of tiling array expression. *Bioinformatics*, 2009.

[59] S. A. Goff, D. Ricke, T.-H. Lan, G. Presting, R. Wang, M. Dunn, J. Glazebrook, A. Sessions, P. Oeller, et al. A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). *Science*, 296 (5565):92–100, 2002.

[60] M. R. Grant, L. Godiard, E. Straube, T. Ashfield, J. Lewald, A. Sattler, R. W. Innes, and J. L. Dangl. Structure of the Arabidopsis RPM1 gene enabling dual specificity disease resistance. *Science*, 269(5225):843–6, 1995.

[61] M. R. Grant, J. M. McDowell, A. G. Sharpe, M. de Torres Zabala, D. J. Lydiate, and J. L. Dangl. Independent deletions of a pathogen-resistance gene in Brassica and Arabidopsis. *Proceedings of the National Academy of Sciences of the USA*, 95(26):15843–8, 1998.

[62] S. Griffiths-Jones, R. J. Grocock, S. van Dongen, A. Bateman, and A. J. Enright. miRBase: MicroRNA sequences, targets and gene nomenclature. *Nucleic Acids Research*, 34(Database issue): D140–4, 2006.

[63] S. S. Gross, C. B. Do, M. Sirota, and S. Batzoglou. CONTRAST: a discriminative, phylogeny-free approach to multiple informant de novo gene prediction. *Genome Biology*, 8(12):R269, 2007.

[64] M. Harbers and P. Carninci. Tag-based approaches for transcriptome research and genome annotation. *Nature Methods*, 2005.

[65] H. He, J. Wang, T. Liu, X. S. Liu, T. Li, Y. Wang, Z. Qian, H. Zheng, X. Zhu, et al. Mapping the C. elegans noncoding transcriptome with a whole-genome tiling microarray. *Genome Research*, 17 (10):1471–7, 2007.

[66] D. Hekstra, A. R. Taussig, M. Magnasco, and F. Naef. Absolute mRNA concentrations from sequence-specific calibration of oligonucleotide arrays. *Nucleic Acids Research*, 31(7):1962–8, 2003.

[67] R. Hettich and K. Kortanek. Semi-infinite programming: Theory, methods, and applications. *SIAM Review*, 1993.

[68] D. A. Hinds, L. L. Stuve, G. B. Nilsen, E. Halperin, E. Eskin, D. G. Ballinger, K. A. Frazer, and D. R. Cox. Whole-genome patterns of common DNA variation in three human populations. *Science*, 307(5712):1072–9, 2005.

[69] D. A. Hinds, A. P. Kloek, M. Jen, X. Chen, and K. A. Frazer. Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nature Genetics*, 38(1):82–5, 2006.

[70] A. Hoerl and R. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(3):55–67, 1970.

[71] I. Hofacker, W. Fontana, P. Stadler, and L. Bonhoeffer. Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie/Chemical Monthly*, 1994.

[72] W. Huber, J. Toedling, and L. M. Steinmetz. Transcript mapping with high-density oligonucleotide tiling arrays. *Bioinformatics*, 22(16):1963–70, 2006.

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

[74] R. A. Irizarry, B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs, and T. P. Speed. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research*, 31(4):e15, 2003.

[75] M. Ishitani, L. Xiong, B. Stevenson, and J. K. Zhu. Genetic analysis of osmotic and cold stress signal transduction in Arabidopsis: Interactions and convergence of abscisic acid-dependent and abscisic acid-independent pathways. *The Plant Cell*, 9(11):1935–49, 1997.

[76] G. Jander, S. R. Norris, S. D. Rounsley, D. F. Bush, I. M. Levin, and R. L. Last. Arabidopsis map-based cloning in the post-genome era. *Plant Physiology*, 129(2):440–50, 2002.

[77] H. Ji and W. H. Wong. TileMap: Create chromosomal map of tiling array hybridizations. *Bioinformatics*, 21(18):3629–36, 2005.

[78] U. Johanson, J. West, C. Lister, S. Michaels, R. Amasino, and C. Dean. Molecular analysis of FRIGIDA, a major determinant of natural variation in Arabidopsis flowering time. *Science*, 290 (5490):344–7, 2000.

[79] J. M. Johnson, S. Edwards, D. Shoemaker, and E. E. Schadt. Dark matter in the genome: Evidence of widespread transcription detected by microarray tiling experiments. *Trends in Genetics*, 21(2): 93–102, 2005.

[80] A. M. Jones, J. Chory, J. L. Dangl, M. Estelle, S. E. Jacobsen, E. M. Meyerowitz, M. Nordborg, and D. Weigel. The impact of Arabidopsis on human health: Diversifying our portfolio. *Cell*, 133 (6):939–43, 2008.

[81] J. D. G. Jones and J. L. Dangl. The plant immune system. *Nature*, 444(7117):323–9, 2006.

[82] M. W. Jones-Rhoades, D. P. Bartel, and B. Bartel. MicroRNAs and their regulatory roles in plants. *Annual Reviews of Plant Biology*, 57:19–53, 2006.

[83] K. Juneau, C. Palm, M. Miranda, and R. W. Davis. High-density yeast-tiling array reveals previously undiscovered introns and extensive regulation of meiotic splicing. *Proceedings of the National Academy of Sciences of the USA*, 104(5):1522–7, 2007.

[84] J. Kaiser. DNA sequencing. a plan to capture human diversity in 1000 genomes. *Science*, 319(5862): 395, 2008.

[85] D. Kampa, J. Cheng, P. Kapranov, M. Yamanaka, S. Brubaker, S. Cawley, J. Drenkow, A. Piccolboni, S. Bekiranov, et al. Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Research*, 14(3):331–42, 2004.

[86] P. Kapranov, S. E. Cawley, J. Drenkow, S. Bekiranov, R. L. Strausberg, S. P. A. Fodor, and T. R. Gingeras. Large-scale transcriptional activity in chromosomes 21 and 22. *Science*, 296(5569):916–9, 2002.

[87] P. Kapranov, J. Cheng, S. Dike, D. A. Nix, R. Duttagupta, A. T. Willingham, P. F. Stadler, J. Hertel, J. Hackermüller, et al. RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science*, 316(5830):1484–8, 2007.

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

[89] J. Kilian, D. Whitehead, J. Horak, D. Wanke, S. Weinl, O. Batistic, C. D'Angelo, E. Bornberg-Bauer, J. Kudla, and K. Harter. The AtGenExpress global stress expression data set: protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses. *The Plant Journal*, 50 (2):347–63, 2007.

[90] S. Kim, V. Plagnol, T. T. Hu, C. Toomajian, R. M. Clark, S. Ossowski, J. R. Ecker, D. Weigel, and M. Nordborg. Recombination and linkage disequilibrium in Arabidopsis thaliana. *Nature Genetics*, 39(9):1151–5, 2007.

[91] J. Korbel, A. Abyzov, X. Mu, N. Carriero, P. Cayting, Z. Zhang, M. Snyder, and M. Gerstein. PEMer: A computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biology*, 10(2):R23, 2009.

[92] J. A. Kreps, Y. Wu, H.-S. Chang, T. Zhu, X. Wang, and J. F. Harper. Transcriptome changes for Arabidopsis in response to salt, osmotic, and cold stress. *Plant Physiology*, 130(4):2129–41, 2002.

[93] Y. Kurihara, A. Matsui, K. Hanada, M. Kawashima, J. Ishida, T. Morosawa, M. Tanaka, E. Kaminuma, Y. Mochizuki, et al. Genome-wide suppression of aberrant mRNA-like noncoding RNAs by NMD in Arabidopsis. *Proceedings of the National Academy of Sciences of the USA*, 106 (7):2453–2458, 2009.

[94] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of the ICML*, 2001.

[95] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, et al., and the International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[96] S. Laubinger, T. Sachsenberg, G. Zeller, W. Busch, J. U. Lohmann, G. Rätsch, and D. Weigel. Dual roles of the nuclear cap-binding complex and SERRATE in pre-mRNA splicing and microRNA processing in Arabidopsis thaliana. *Proceedings of the National Academy of Sciences of the USA*, 105(25):8795–800, 2008.

[97] S. Laubinger, G. Zeller, S. R. Henz, T. Sachsenberg, C. K. Widmer, N. Naouar, M. Vuylsteke, B. Schölkopf, G. Rätsch, and D. Weigel. At-TAX: A whole genome tiling array resource for developmental expression analysis and transcript identification in Arabidopsis thaliana. *Genome Biology*, 9 (7):R112, 2008.

[98] S. Laubinger, G. Zeller, S. R. Henz, T. Sachsenberg, G. Rätsch, and D. Weigel. Global effects of the small RNA biogenesis machinery on the Arabidopsis transcriptome. *Manuscript in preparation*, 2009.

[99] I. Lee, A. A. Dombkowski, and B. D. Athey. Guidelines for incorporating non-perfectly matched oligonucleotides into target-specific hybridization probes for a DNA microarray. *Nucleic Acids Research*, 32(2):681–90, 2004.

[100] J.-Y. Lee, J. Colinas, J. Y. Wang, D. Mace, U. Ohler, and P. N. Benfey. Transcriptional and posttranscriptional regulation of transcription factor expression in Arabidopsis roots. *Proceedings of the National Academy of Sciences of the USA*, 103(15):6055–60, 2006.

[101] S. Levy, G. Sutton, P. C. Ng, L. Feuk, A. L. Halpern, B. P. Walenz, N. Axelrod, J. Huang, E. F. Kirkness, et al. The diploid genome sequence of an individual human. *PLoS Biology*, 5(10):e254, 2007.

[102] L. Li, X. Wang, V. Stolc, X. Li, D. Zhang, N. Su, W. Tongprasit, S. Li, Z. Cheng, et al. Genome-wide transcription analyses in rice using tiling microarrays. *Nature Genetics*, 38(1):124–9, 2006.

[103] L. Li, X. Wang, R. Sasidharan, V. Stolc, W. Deng, H. He, J. Korbel, X. Chen, W. Tongprasit, et al. Global identification and characterization of transcriptionally active regions in the rice genome. *PLoS ONE*, 2(3):e294, 2007.

[104] W. Li, C. A. Meyer, and X. S. Liu. A hidden Markov model for analyzing ChIP-chip experiments on genome tiling arrays and its application to p53 binding sequences. *Bioinformatics*, 21 Suppl 1: i274–82, 2005.

[105] S. E. V. Linsen, E. de Wit, G. Janssens, S. Heater, L. Chapman, R. K. Parkin, B. Fritz, S. K. Wyman, E. de Bruijn, et al. Limitations and possibilities of small RNA digital gene expression profiling. *Nature Methods*, 6(7):474–6, 2009.

[106] R. Lister, R. C. O'Malley, J. Tonti-Filippini, B. D. Gregory, C. C. Berry, A. H. Millar, and J. R. Ecker. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell*, 133(3): 523–36, 2008.

[107] H.-H. Liu, X. Tian, Y.-J. Li, C.-A. Wu, and C.-C. Zheng. Microarray-based analysis of stress-regulated microRNAs in Arabidopsis thaliana. *RNA*, 14(5):836–43, 2008.

[108] J. R. Manak, S. Dike, V. Sementchenko, P. Kapranov, F. Biemar, J. Long, J. Cheng, I. Bell, S. Ghosh, et al. Biological function of unannotated transcription during the early development of Drosophila melanogaster. *Nature Genetics*, 38(10):1151–8, 2006.

[109] T. P. Mann and W. S. Noble. Efficient identification of DNA hybridization partners in a sequence database. *Bioinformatics*, 22(14):e350–8, 2006.

[110] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, 18(9): 1509–17, 2008.

[111] K. McNally, K. Childs, R. Bohnert, R. Davidson, K. Zhao, V. Ulat, R. Clark, G. Zeller, D. Hoen, et al. Genome-wide SNP variation reveals relationships among landraces and modern varieties of rice. *Proceedings of the National Academy of Sciences of the USA*, 106:(30):12273–78, 2009.

[112] K. L. McNally, R. Bruskiewich, D. Mackill, C. R. Buell, J. E. Leach, and H. Leung. Sequencing multiple and diverse rice varieties. connecting whole-genome variation with phenotypes. *Plant Physiology*, 141(1):26–31, 2006.

[113] R. Mei, E. Hubbell, S. Bekiranov, M. Mittmann, F. C. Christians, M.-M. Shen, G. Lu, J. Fang, W.-M. Liu, et al. Probe selection for high-density oligonucleotide arrays. *Proceedings of the National Academy of Sciences of the USA*, 100(20):11237–42, 2003.

[114] B. C. Meyers, S. S. Tej, T. H. Vu, C. D. Haudenschild, V. Agrawal, S. B. Edberg, H. Ghazal, and S. Decola. The use of MPSS for whole-genome transcriptional analysis in Arabidopsis. *Genome Research*, 14(8):1641–53, 2004.

[115] B. C. Meyers, T. H. Vu, S. S. Tej, H. Ghazal, M. Matvienko, V. Agrawal, J. Ning, and C. D. Haudenschild. Analysis of the transcriptional complexity of Arabidopsis thaliana by massively parallel signature sequencing. *Nature Biotechnology*, 22(8):1006–11, 2004.

[116] R. E. Mills, C. T. Luttig, C. E. Larkins, A. Beauchamp, C. Tsui, W. S. Pittard, and S. E. Devine. An initial map of insertion and deletion (indel) variation in the human genome. *Genome Research*, 16(9):1182–90, 2006.

[117] T. C. Mockler, S. Chan, A. Sundaresan, H. Chen, S. E. Jacobsen, and J. R. Ecker. Applications of DNA tiling arrays for whole-genome analysis. *Genomics*, 85(1):1–15, 2005.

[118] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nature Methods*, 5(7):621–8, 2008.

[119] L. J. Mullan and A. J. Bleasby. Short EMBOSS user guide. European molecular biology open software suite. *Briefings in Bioinformatics*, 3(1):92–4, 2002.

[120] K. Munch, P. P. Gardner, P. Arctander, and A. Krogh. A hidden Markov model approach for determining expression from genomic tiling micro arrays. *BMC Bioinformatics*, 7:239, 2006.

[121] R. J. Mural, M. D. Adams, E. W. Myers, H. O. Smith, G. L. G. Miklos, R. Wides, A. Halpern, P. W. Li, G. G. Sutton, et al. A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. *Science*, 296(5573):1661–71, 2002.

[122] F. Naef and M. O. Magnasco. Solving the riddle of the bright mismatches: Labeling and effective binding in oligonucleotide arrays. *Physical review E, Statistical, nonlinear, and soft matter physics*, 68(1 Pt 1):011906, 2003.

[123] N. Naouar, K. Vandepoele, T. Lammens, T. Casneuf, G. Zeller, P. van Hummelen, D. Weigel, G. Rätsch, D. Inzé, et al. Quantitative RNA expression analysis with Affymetrix tiling 1.0r arrays identifies new E2F target genes. *The Plant Journal*, 2008.

[124] T. Nawy, J.-Y. Lee, J. Colinas, J. Y. Wang, S. C. Thongrod, J. E. Malamy, K. Birnbaum, and P. N. Benfey. Transcriptional profile of the Arabidopsis root quiescent center. *The Plant Cell*, 17 (7):1908–25, 2005.

[125] A. Ng and M. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. *Proceedings of NIPS*, 2002.

[126] N. Nguyen and Y. Guo. Comparisons of sequence labeling algorithms and extensions. *Proceedings of the ICML*, 2007.

[127] M. Nordborg and D. Weigel. Next-generation genetics in plants. *Nature*, 456(7223):720–3, 2008.

[128] M. Nordborg, T. T. Hu, Y. Ishino, J. Jhaveri, C. Toomajian, H. Zheng, E. Bakker, P. Calabrese, J. Gladstone, et al. The pattern of polymorphism in Arabidopsis thaliana. *PLoS Biology*, 3(7):e196, 2005.

[129] T. R. O'Connor, C. Dyreson, and J. J. Wyrick. Athena: A resource for rapid visualization and systematic analysis of Arabidopsis promoter sequences. *Bioinformatics*, 21(24):4411–3, 2005.

[130] S. Ossowski, K. Schneeberger, R. Clark, C. Lanz, N. Warthmann, and D. Weigel. Sequencing of natural strains of Arabidopsis thaliana with short reads. *Genome Research*, 2008.

[131] N. Patil, A. Berno, D. Hinds, W. Barrett, and J. Doshi. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, 2001.

[132] A. C. Pease, D. Solas, E. J. Sullivan, M. T. Cronin, C. P. Holmes, and S. P. Fodor. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proceedings of the National Academy of Sciences of the USA*, 91(11):5022–6, 1994.

[133] A. Piccolboni. Multivariate segmentation in the analysis of transcription tiling array data. *Lecture Notes in Computer Science*, 2007.

[134] D. Pinkel, R. Segraves, D. Sudar, S. Clark, I. Poole, D. Kowbel, C. Collins, W. L. Kuo, C. Chen, et al. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics*, 20(2):207–11, 1998.

[135] R. Rajagopalan, H. Vaucheret, J. Trejo, and D. P. Bartel. A diverse and evolutionarily fluid set of microRNAs in Arabidopsis thaliana. *Genes & Development*, 20(24):3407–25, 2006.

[136] G. Rätsch and S. Sonnenburg. Accurate splice site detection for Caenorhabditis elegans. In *Kernel Methods in Computational Biology*, pages 277–298. MIT Press, 2004.

[137] G. Rätsch and S. Sonnenburg. Large scale hidden semi-Markov SVMs. *Proceedings of NIPS*, 2007.

[138] G. Rätsch, A. Demiriz, and K. Bennett. Sparse regression ensembles in infinite and finite hypothesis spaces. *Machine Learning*, 2002.

[139] G. Rätsch, S. Sonnenburg, and B. Schölkopf. RASE: Recognition of alternatively spliced exons in C. elegans. *Bioinformatics*, 21 Suppl 1:i369–77, 2005.

[140] G. Rätsch, S. Sonnenburg, J. Srinivasan, H. Witte, K.-R. Müller, R.-J. Sommer, and B. Schölkopf. Improving the Caenorhabditis elegans genome annotation using machine learning. *PLoS Computational Biology*, 3(2):e20, 2007.

[141] J. C. Redman, B. J. Haas, G. Tanimoto, and C. D. Town. Development and evaluation of an Arabidopsis whole genome Affymetrix probe array. *The Plant Journal*, 38(3):545–61, 2004.

[142] P. Rice, I. Longden, and A. Bleasby. EMBOSS: the European molecular biology open software suite. *Trends in Genetics*, 16(6):276–7, 2000.

[143] J. L. Rinn, G. Euskirchen, P. Bertone, R. Martone, N. M. Luscombe, S. Hartman, P. M. Harrison, F. K. Nelson, P. Miller, et al. The transcriptional activity of human chromosome 22. *Genes & Development*, 17(4):529–40, 2003.

[144] T. E. Royce, J. S. Rozowsky, P. Bertone, M. Samanta, V. Stolc, S. Weissman, M. Snyder, and M. Gerstein. Issues in the analysis of oligonucleotide tiling microarrays for transcript mapping. *Trends in Genetics*, 21(8):466–75, 2005.

[145] T. E. Royce, J. S. Rozowsky, and M. B. Gerstein. Assessing the need for sequence-based normalization in tiling microarray experiments. *Bioinformatics*, 23(8):988–97, 2007.

[146] S. Rozen and H. Skaletsky. Primer3 on the www for general users and for biologist programmers. *Methods in Molecular Biology*, 132:365–86, 2000.

[147] M. P. Samanta, W. Tongprasit, H. Sethi, C.-S. Chin, and V. Stolc. Global identification of noncoding RNAs in Saccharomyces cerevisiae by modulating an essential RNA processing pathway. *Proceedings of the National Academy of Sciences of the USA*, 103(11):4192–7, 2006.

[148] J. Schacherer, J. A. Shapiro, D. M. Ruderfer, and L. Kruglyak. Comprehensive polymorphism survey elucidates population structure of Saccharomyces cerevisiae. *Nature*, 458(7236):342–5, 2009.

[149] E. E. Schadt, S. W. Edwards, D. GuhaThakurta, D. Holder, L. Ying, V. Svetnik, A. Leonardson, K. W. Hart, A. Russell, et al. A comprehensive transcript index of the human genome generated using microarrays and computational approaches. *Genome Biology*, 5(10):R73, 2004.

[150] K. J. Schmid, T. R. Sorensen, R. Stracke, O. Torjek, T. Altmann, T. Mitchell-Olds, and B. Weisshaar. Large-scale identification and analysis of genome-wide single-nucleotide polymorphisms for mapping in Arabidopsis thaliana. *Genome Research*, 13(6A):1250–7, 2003.

[151] K. J. Schmid, S. Ramos-Onsins, H. Ringys-Beckstein, B. Weisshaar, and T. Mitchell-Olds. A multi-locus sequence survey in Arabidopsis thaliana reveals a genome-wide departure from a neutral model of DNA sequence polymorphism. *Genetics*, 169(3):1601–15, 2005.

[152] M. Schmid, T. S. Davison, S. R. Henz, U. J. Pape, M. Demar, M. Vingron, B. Schölkopf, D. Weigel, and J. U. Lohmann. A gene expression map of Arabidopsis thaliana development. *Nature Genetics*, 37(5):501–6, 2005.

[153] B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, 2002.

[154] U. Schulze, B. Hepp, C. S. Ong, and G. Rätsch. PALMA: mRNA to genome alignments using large margin algorithms. *Bioinformatics*, 23(15):1892–900, 2007.

[155] G. Schweikert, A. Zien, G. Zeller, J. Behr, C. Dieterich, C. Ong, P. Philips, F. De Bona, L. Hartmann, et al. mGene: Accurate SVM-based gene finding with an application to nematode genomes. *Genome Research*, 2009.

[156] M. Seki, M. Narusaka, H. Abe, M. Kasuga, K. Yamaguchi-Shinozaki, P. Carninci, Y. Hayashizaki, and K. Shinozaki. Monitoring the expression pattern of 1300 Arabidopsis genes under drought and cold stresses by using a full-length cDNA microarray. *The Plant Cell*, 13(1):61–72, 2001.

[157] M. Seki, M. Narusaka, A. Kamiya, J. Ishida, M. Satou, T. Sakurai, M. Nakajima, A. Enju, K. Akiyama, et al. Functional annotation of a full-length Arabidopsis cDNA collection. *Science*, 296 (5565):141–5, 2002.

[158] D. W. Selinger, K. J. Cheung, R. Mei, E. M. Johansson, C. S. Richmond, F. R. Blattner, D. J. Lockhart, and G. M. Church. RNA expression analysis using a 30 base pair resolution Escherichia coli genome array. *Nature Biotechnology*, 18(12):1262–8, 2000.

[159] F. Sha and L. Saul. Large margin hidden Markov models for automatic speech recognition. *Proceedings of NIPS*, 2007.

[160] J. Shen, H. Araki, L. Chen, J.-Q. Chen, and D. Tian. Unique evolutionary mechanism in R-genes under the presence/absence polymorphism in Arabidopsis thaliana. *Genetics*, 172(2):1243–50, 2006.

[161] J. Shendure. The beginning of the end for microarrays? *Nature Methods*, 5(7):585–7, 2008.

[162] J. Shendure, R. D. Mitra, C. Varma, and G. M. Church. Advanced sequencing technologies: Methods and goals. *Nature Reviews Genetics*, 5(5):335–44, 2004.

[163] S. Sherman-Broyles, N. Boggs, A. Farkas, P. Liu, J. Vrebalov, M. E. Nasrallah, and J. B. Nasrallah. S locus genes and the evolution of self-fertility in Arabidopsis thaliana. *The Plant Cell*, 19(1):94–106, 2007.

[164] M. Shinawi and S. W. Cheung. The array cgh and its clinical applications. *Drug Discovery Today*, 13(17-18):760–70, 2008.

[165] D. D. Shoemaker, E. E. Schadt, C. D. Armour, Y. D. He, P. Garrett-Engele, P. D. McDonagh, P. M. Loerch, A. Leonardson, P. Y. Lum, et al. Experimental annotation of the human genome using microarray technology. *Nature*, 409(6822):922–7, 2001.

[166] W. Y. Song, G. L. Wang, L. L. Chen, H. S. Kim, L. Y. Pi, T. Holsten, J. Gardner, B. Wang, W. X. Zhai, et al. A receptor kinase-like protein encoded by the rice disease resistance gene, Xa21. *Science*, 270(5243):1804–6, 1995.

[167] S. Sonnenburg, G. Rätsch, and B. Schölkopf. Large scale genomic sequence SVM classifiers. *Proceedings of the ICML*, 2005.

[168] S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf. Large scale multiple kernel learning. *Journal of Machine Learning Research*, 2006.

[169] S. Sonnenburg, A. Zien, and G. Rätsch. ARTS: Accurate recognition of transcription starts in human. *Bioinformatics*, 22(14):e472–80, 2006.

[170] S. Sonnenburg, G. Schweikert, P. Philips, J. Behr, and G. Rätsch. Accurate splice site prediction using support vector machines. *BMC Bioinformatics*, 8 Suppl 10:S7, 2007.

[171] M. W. B. Spencer, S. A. Casson, and K. Lindsey. Transcriptional profiling of the Arabidopsis embryo. *Plant Physiology*, 143(2):924–40, 2007.

[172] M. Stanke and S. Waack. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*, 19 Suppl 2:ii215–25, 2003.

[173] L. D. Stein, C. Mungall, S. Shu, M. Caudy, M. Mangone, A. Day, E. Nickerson, J. E. Stajich, T. W. Harris, et al. The generic genome browser: A building block for a model organism system database. *Genome Research*, 12(10):1599–610, 2002.

[174] V. Stolc, Z. Gauhar, C. Mason, G. Halasz, M. F. van Batenburg, S. A. Rifkin, S. Hua, T. Herreman, W. Tongprasit, et al. A gene expression map for the euchromatic genome of Drosophila melanogaster. *Science*, 306(5696):655–60, 2004.

[175] V. Stolc, M. P. Samanta, W. Tongprasit, H. Sethi, S. Liang, D. C. Nelson, A. Hegeman, C. Nelson, D. Rancour, et al. Identification of transcribed sequences in Arabidopsis thaliana by using high-resolution genome tiling arrays. *Proceedings of the National Academy of Sciences of the USA*, 102 (12):4453–8, 2005.

[176] M. Suárez-Fariñas, M. Pellegrino, K. M. Wittkowski, and M. O. Magnasco. Harshlight: A corrective make-upprogram for microarray chips. *BMC Bioinformatics*, 6:294, 2005.

[177] R. Sunkar and J.-K. Zhu. Novel and stress-regulated microRNAs and other small RNAs from Arabidopsis. *The Plant Cell*, 16(8):2001–19, 2004.

[178] D. Swarbreck, C. Wilks, P. Lamesch, T. Z. Berardini, M. Garcia-Hernandez, H. Foerster, D. Li, T. Meyer, R. Muller, et al. The Arabidopsis information resource (TAIR): Gene structure and function annotation. *Nucleic Acids Research*, 36(Database issue):D1009–14, 2008.

[179] C. Tang, C. Toomajian, S. Sherman-Broyles, V. Plagnol, Y.-L. Guo, T. T. Hu, R. M. Clark, J. B. Nasrallah, D. Weigel, and M. Nordborg. The evolution of selfing in Arabidopsis thaliana. *Science*, 317(5841):1070–2, 2007.

[180] B. Taskar, C. Guestrin, and D. Koller. Max-margin Markov networks. *Proceedings of NIPS*, 2004.

[181] J. H. Thomas. Adaptive evolution in two large families of ubiquitin-ligase adapters in nematodes and plants. *Genome Research*, 16(8):1017–30, 2006.

[182] C. Toomajian, T. T. Hu, M. J. Aranzana, C. Lister, C. Tang, H. Zheng, K. Zhao, P. Calabrese, C. Dean, and M. Nordborg. A nonparametric test reveals selection for rapid flowering in the Arabidopsis genome. *PLoS Biology*, 4(5):e137, 2006.

[183] T. Toyoda and K. Shinozaki. Tiling array-driven elucidation of transcriptional structures based on maximum-likelihood and Markov models. *The Plant Journal*, 43(4):611–21, 2005.

[184] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 2006.

[185] G. A. Tuskan, S. Difazio, S. Jansson, J. Bohlmann, I. Grigoriev, U. Hellsten, N. Putnam, S. Ralph, S. Rombauts, et al. The genome of black cottonwood, Populus trichocarpa. *Science*, 313(5793): 1596–604, 2006.

[186] V. N. Vapnik. *The nature of statistical learning theory.* Springer Verlag, 1995.

[187] V. E. Velculescu, L. Zhang, B. Vogelstein, and K. W. Kinzler. Serial analysis of gene expression. *Science*, 270(5235):484–7, 1995.

[188] Z. Wang, M. Gerstein, and M. Snyder. RNA-seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009.

[189] D. Weigel and R. Mott. The 1001 genomes project for Arabidopsis thaliana. *Genome Biology*, 10 (107), 2009.

[190] S. R. Wicks, R. T. Yeh, W. R. Gish, R. H. Waterston, and R. H. Plasterk. Rapid gene mapping in Caenorhabditis elegans using a high density polymorphism map. *Nature Genetics*, 28(2):160–4, 2001.

[191] A. J. Windsor, M. E. Schranz, N. Formanová, S. Gebauer-Jung, J. G. Bishop, D. Schnabelrauch, J. Kroymann, and T. Mitchell-Olds. Partial shotgun sequencing of the Boechera stricta genome reveals extensive microsynteny and promoter conservation with Arabidopsis. *Plant Physiology*, 140 (4):1169–82, 2006.

[192] S. I. Wright and B. S. Gaut. Molecular population genetics and the search for adaptive evolution in plants. *Molecular Biology and Evolution*, 22(3):506–19, 2005.

[193] S. C. Wu, J. Györgyey, and D. Dudits. Polyadenylated H3 histone transcripts and H3 histone variants in alfalfa. *Nucleic Acids Research*, 17(8):3057–63, 1989.

[194] Z. Wu and R. A. Irizarry. Stochastic models inspired by hybridization theory for short oligonucleotide arrays. *Journal of Computational Biology*, 12(6):882–93, 2005.

[195] Z. Wu, R. Irizarry, R. Gentleman, and F. M.-M. .... A model-based background adjustment for oligonucleotide expression arrays. *Journal of the American Statistical Association*, 2004.

[196] Z. Xie, E. Allen, N. Fahlgren, A. Calamar, S. A. Givan, and J. C. Carrington. Expression of Arabidopsis MIRNA genes. *Plant Physiology*, 138(4):2145–54, 2005.

[197] L. Xiong, K. S. Schumaker, and J.-K. Zhu. Cell signaling during cold, drought, and salt stress. *The Plant Cell*, 14 Suppl:S165–83, 2002.

[198] K. Yamada, J. Lim, J. M. Dale, H. Chen, P. Shinn, C. J. Palm, A. M. Southwick, H. C. Wu, C. Kim, et al. Empirical analysis of transcriptional activity in the Arabidopsis genome. *Science*, 302(5646): 842–6, 2003.

[199] K. Yamaguchi-Shinozaki and K. Shinozaki. Transcriptional regulatory networks in cellular responses and tolerance to dehydration and cold stresses. *Annual Reviews of Plant Biology*, 57:781–803, 2006.

[200] J. Yazaki, B. D. Gregory, and J. R. Ecker. Mapping the genome landscape using tiling array technology. *Current Opinion in Plant Biology*, 10(5):534–42, 2007.

[201] J. Yu, S. Hu, J. Wang, G. K.-S. Wong, S. Li, B. Liu, Y. Deng, L. Dai, Y. Zhou, et al. A draft sequence of the rice genome (Oryza sativa L. ssp. indica). *Science*, 296(5565):79–92, 2002.

[202] G. Zeller. Resequencing data of twenty Arabidopsis ecotypes. *Diplom Thesis at Tübingen University*, 2005.

[203] G. Zeller, R. M. Clark, K. Schneeberger, A. Bohlen, D. Weigel, and G. Rätsch. Detecting polymorphic regions in Arabidopsis thaliana with resequencing microarrays. *Genome Research*, 18(6):918–29, 2008.

[204] G. Zeller, S. R. Henz, S. Laubinger, D. Weigel, and G. Rätsch. Transcript normalization and segmentation of tiling array data. *Pacific Symposium on Biocomputing*, pages 527–38, 2008.

[205] G. Zeller, S. Henz, C. Widmer, T. Sachsenberg, G. Rätsch, D. Weigel, and S. Laubinger. Stress-induced changes in the Arabidopsis thaliana transcriptome analyzed using whole genome tiling arrays. *The Plant Journal*, 2009.

[206] Y. Zhan and D. Kulp. Model-P: A basecalling method for resequencing microarrays of diploid samples. *Bioinformatics*, 21 Suppl 2:ii182–9, 2005.

[207] L. Zhang, M. F. Miles, and K. D. Aldape. A model of molecular interactions on short oligonucleotide microarrays. *Nature Biotechnology*, 21(7):818–21, 2003.

[208] X. Zhang, J. Yazaki, A. Sundaresan, S. Cokus, S. W.-L. Chan, H. Chen, I. R. Henderson, P. Shinn, M. Pellegrini, et al. Genome-wide high-resolution mapping and functional analysis of DNA methylation in Arabidopsis. *Cell*, 126(6):1189–201, 2006.

[209] J.-K. Zhu. Salt and drought stress signal transduction in plants. *Annual Reviews of Plant Biology*, 53:247–73, 2002.

Appendix

Supplementary Protocols

Hybridization protocol for Arabidopsis tiling arrays

For synthesis of hybridization targets, 1 μg of total RNA was used as template for generation of cRNA using the MessageAmp II-Biotin Enhanced Kit (Ambion, Austin, TX, USA). We followed the manufacturer's instruction with one exception: Biotinylated NTPs were replaced by unmodified NTPs (stock solution 25 mM each). 7 μg unmodified cRNA was converted into dsDNA (GeneChip© WT Double-Stranded cDNA Synthesis Kit, Affymetrix, Santa Clara, CA, USA) and dsDNA was purified using the MinElute Reaction Cleanup Kit (Qiagen, Hilden, Germany). 7.5 μg dsDNA was fragmented and labeled using the GeneChip© WT Double-Stranded DNA Terminal Labeling Kit (Affymetrix). Targets were hybridized to *Arabidopsis* Tiling 1.0R arrays for 14 h at 42°C, washed (Fluidics Station 450, wash protocol FS450_0001) and scanned using a GeneChip© Scanner 3000 7G. For comparison of polyA(+) and polyA(+/-), rRNA was depleted from 10 μg total RNA using RiboMinus™ Yeast Transcriptome Isolation Kit (Invitrogen) and an *Arabidopsis*-specific RiboMinus™ LNA oligonucleotide mix kindly provided by Invitrogen. rRNA-depleted RNA was precipitated and resuspended in 12 μl water, from which 11 μl were used for reverse transcription using MessageAmp II-Biotin Enhanced Kit (Ambion) with an oligo-dT-T7 primer (MessageAmp II-Biotin Enhanced Kit) or a random-T7 primer (included in the GeneChip© WT Amplified Double-Stranded cDNA Synthesis Kit, Affymetrix).

Supplementary Tables

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

Table 4.1: Whole-genome annotation with repetitive resequencing probes for *A. thaliana*. [a] Pairs of genomic positions with similar probe sequence by match type criteria. [b] Unique positions tiled on the arrays corresponding to the various repetitive classes.

Accession	Number of PRs	% genome in PRs	Precision	Recall
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 4.2: 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 3.1, $\lambda = 75\%$). Precision and recall for each accession as determined from 2010 is also given. 2010 data for Van-0 was not available; nevertheless, we used HM-SVMs trained across data from all other accessions to predict PRs in Van-0. The absence of test data precluded evaluation of precision and recall for the Van-0 accession (NA is "not applicable").

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

Series	Sample-ID	Replicates	Tissue	Genotype	Amplification	Growth conditions
D	D_001	3 biological	roots	Col-0	oligo-dT	
D	D_002	3 biological	seedlings	Col-0	oligo-dT	
D	D_003	3 biological	young leaves	Col-0	oligo-dT	
D	D_004	3 biological	senescing leaves	Col-0	oligo-dT	
D	D_005	3 biological	vegetative apex	Col-0	oligo-dT	
D	D_006	3 biological	inflorescence apex	Col-0	oligo-dT	
D	D_007	3 biological	stem	Col-0	oligo-dT	
D	D_008	3 biological	inflorescences	Col-0	oligo-dT	
D	D_009	3 biological	inflorescences	*clv-3*	oligo-dT	
D	D_010	3 biological	flowers	Col-0	oligo-dT	
D	D_011	3 biological	fruits	Col-0	oligo-dT	
D	D_012	3 biological	inflorescences	Col-0	oligo-dT	
D	D_013	3 biological	seedlings	Col-0	oligo-dT	
D	D_014	3 biological	inflorescences	Col-0	random	
D	D_015	3 biological	seedlings	Col-0	random	
S	S_001	3 biological	seedlings	Col-0	oligo-dT	control at t=0
S	S_002	3 biological	seedlings	Col-0	oligo-dT	mock control at t=1 h
S	S_003	3 biological	seedlings	Col-0	oligo-dT	mock control at t=12 h
S	S_004	3 biological	seedlings	Col-0	oligo-dT	1 h salt stress
S	S_005	3 biological	seedlings	Col-0	oligo-dT	12 h salt stress
S	S_006	3 biological	seedlings	Col-0	oligo-dT	1 h osmostic stress
S	S_007	3 biological	seedlings	Col-0	oligo-dT	12 h osmotic stress
S	S_008	3 biological	seedlings	Col-0	oligo-dT	1 h ABA^a treatment
S	S_009	3 biological	seedlings	Col-0	oligo-dT	12 h ABA^a treatment
S	S_010	3 biological	seedlings	Col-0	oligo-dT	1 h cold stress (10°C)
S	S_011	3 biological	seedlings	Col-0	oligo-dT	12 h cold stress (10°C)
S	S_012	3 biological	seedlings	Col-0	oligo-dT	1 h heat stress (30°C)
S	S_013	3 biological	seedlings	Col-0	oligo-dT	12 h heat stress (30°C)
T	T_003	3 technical	inflorescences	Col-0	oligo-dT	
T	T_004	3 technical	roots	Col-0	oligo-dT	

Table 4.4: Plant samples used for tiling array transcriptome analyses. Arrays of the same series were hybridized with RNA from plants grown under comparable conditions and data was normalized and analyzed together. (a) Abscisic acid (ABA), is a plant hormone that plays a major role in mediating stress response [reviewed in 49, 209]. Data from the D series was analyzed in Laubinger et al. [97]. Zeller et al. [205] was based on data from the S series. Array data from the T series was used for experiments in Zeller et al. [204].

Supplementary Figures

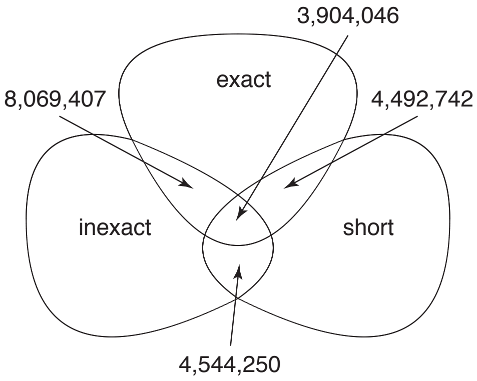


Figure 4.1: Intersection between non-redundant positions tiled on resequencing arrays with 25mer matches. For example, of 8,069,407 positions where there is an exact and inexact 25mer match, 3,904,046 also have a short 25mer match. Absolute numbers for match types are given in Table 4.1

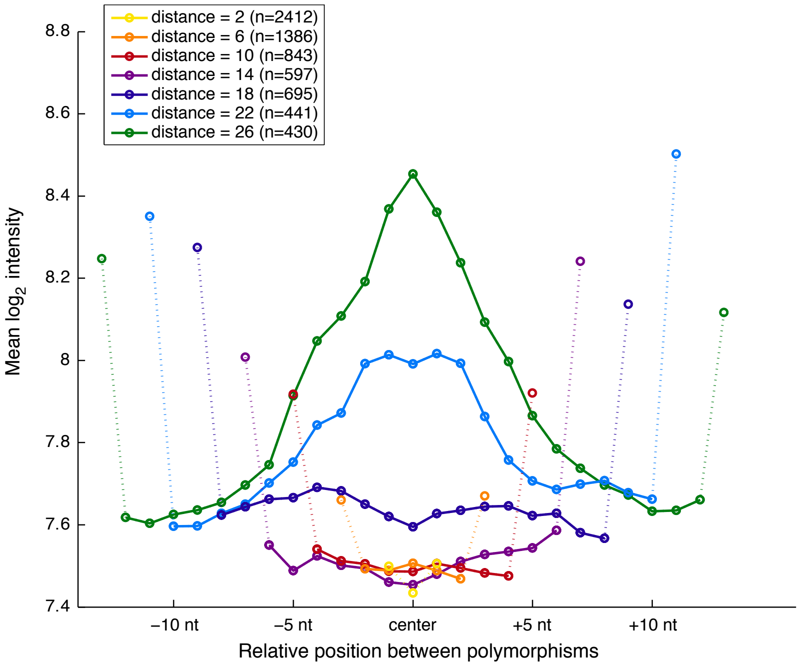


Figure 4.2: Intensities for features located between adjacent polymorphisms is reduced. Regions between polymorphisms at a distance ≤26 bp to each other were extracted and categorized according to this distance (see inset). For each distance category the maximal intensities for each probe quartet between 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).

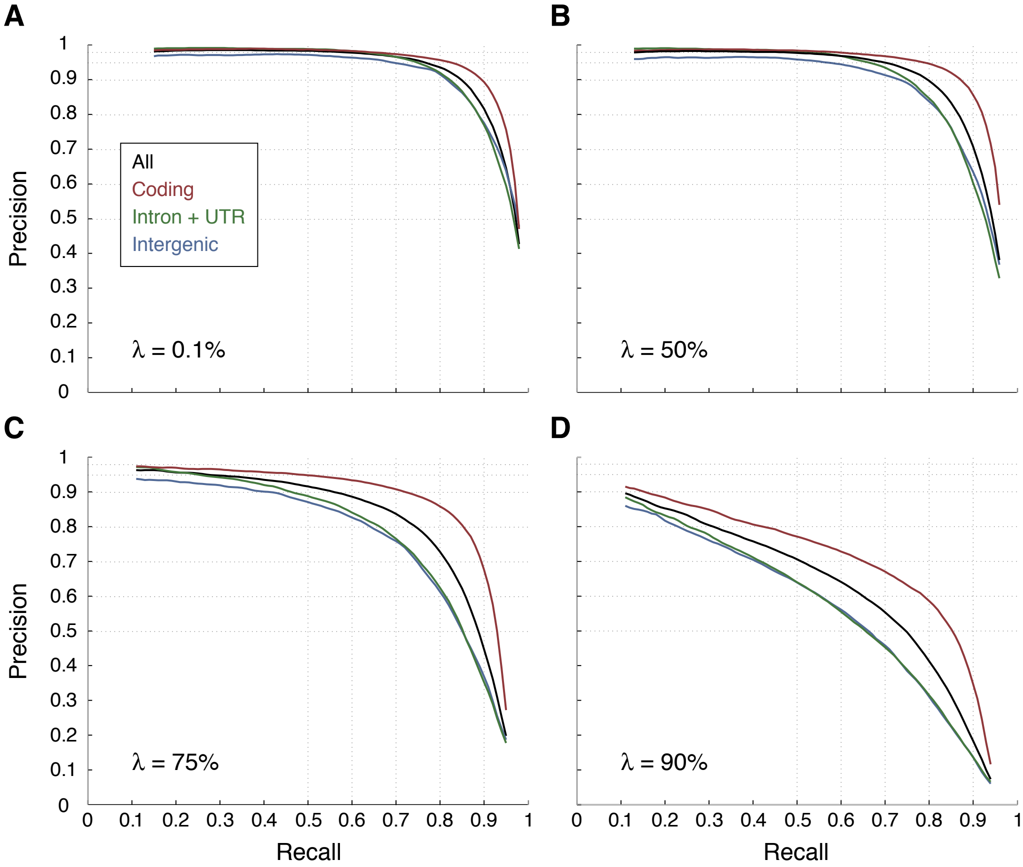


Figure 4.3: Dependency of performance on the choice of the minimal required overlap λ between known PRs and PR predictions. Shown are the precision-recall curves for 4 different choices of λ (panels **A**-**D**).

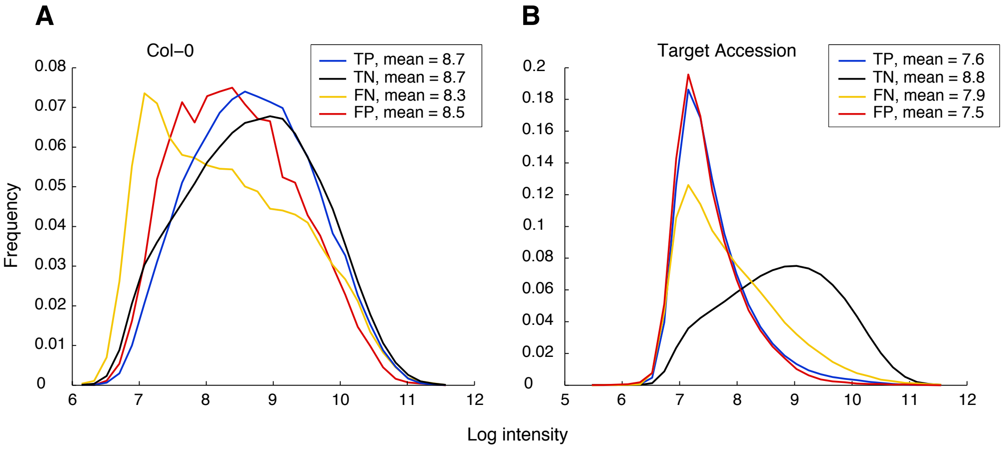


Figure 4.4: Detection of PRs is difficult in regions where hybridization intensities for the reference accession, Col-0, were reduced due to unfavourable hybridization properties of oligonucleotides. Intensity histograms were calculated separately for true positive (TP), false positive (FP), true negative (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.

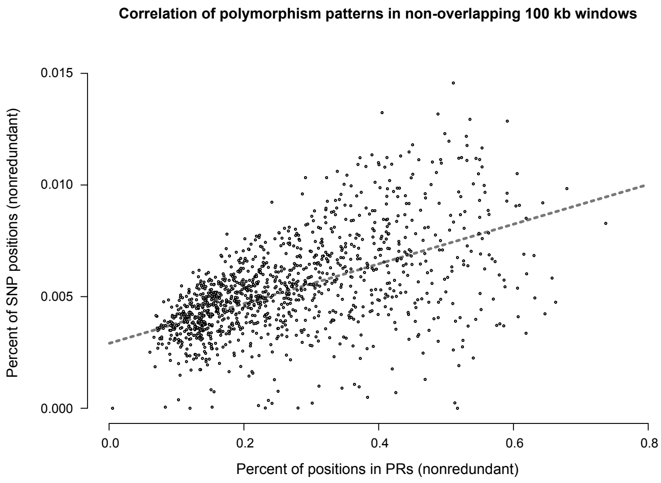


Figure 4.5: Correlation between estimates of polymorphism from PR predictions and MBML2 SNPs. Polymorphism was calculated as in Fig. 3.8 for positions central to non-overlapping 100 kb windows, and estimates from the two data sets are significantly correlated (Pearson's cor = 0.54, P-value ¡ 10^{-15}), even though the estimates sometimes differ substantially (see also Fig. 3.8). 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 3.3). 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).

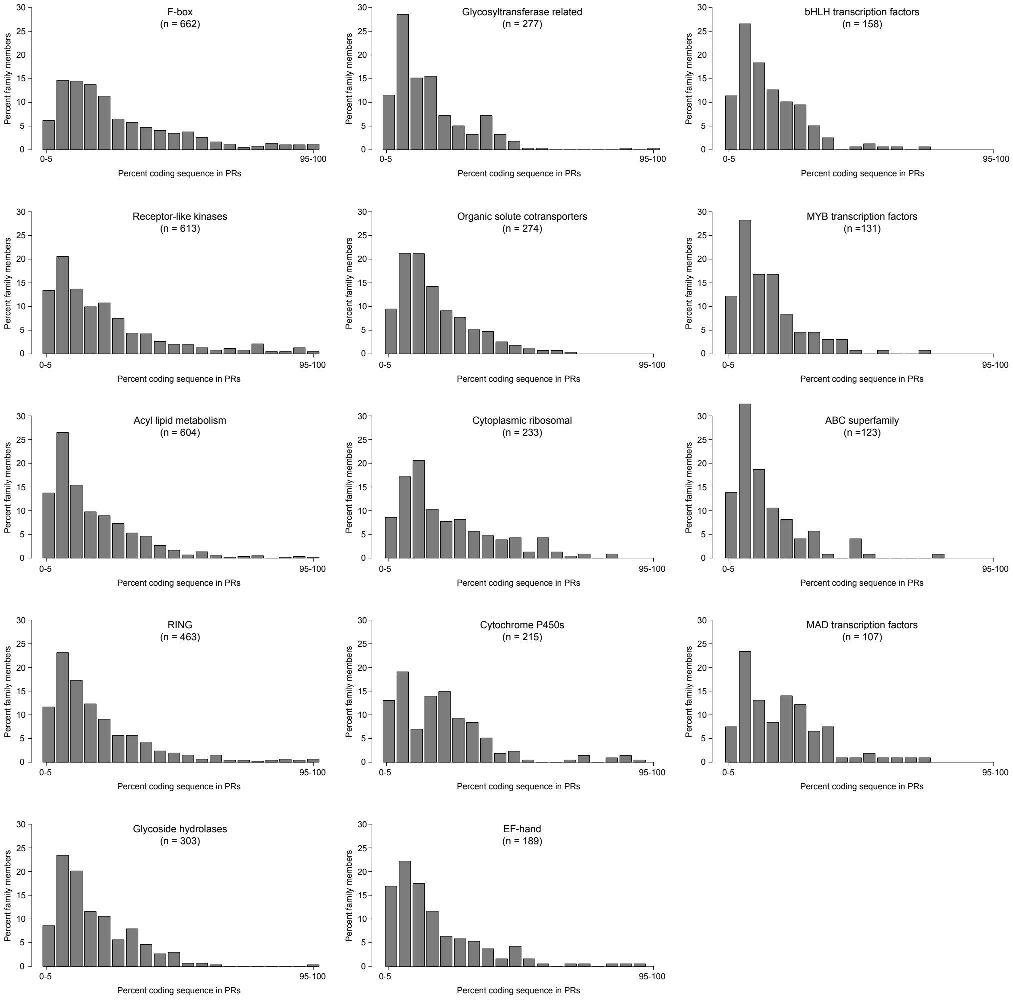


Figure 4.6: Distribution of coding genes by percent inclusion in PRs by gene family classification. See Fig. 3.10 A, B for additional information and gene families.

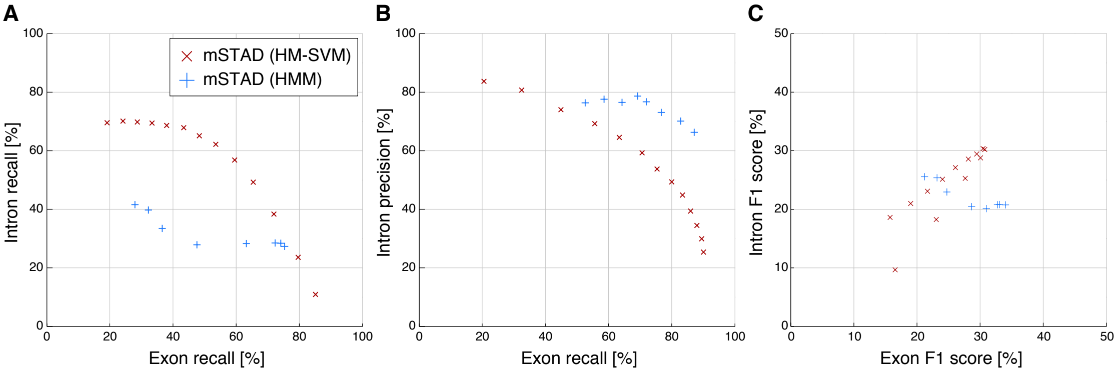


Figure 4.7: mSTAD HM-SVM and mSTAD HMM differ in the way exon prediction accuracy is traded off against intron accuracy. (see inset, Section 3.3.2, and Fig. 3.19 for definition of precision and recall on exon and intron level). **(A)** Balance between exon and intron recall. **(B)** Balance between exon and intron precision. **(C)** Balance between F1 scores for exons and introns. The F1 score is the harmonic mean between precision and recall: $F1 = (\text{precision} \times \text{recall})/(\text{precision} + \text{recall})$

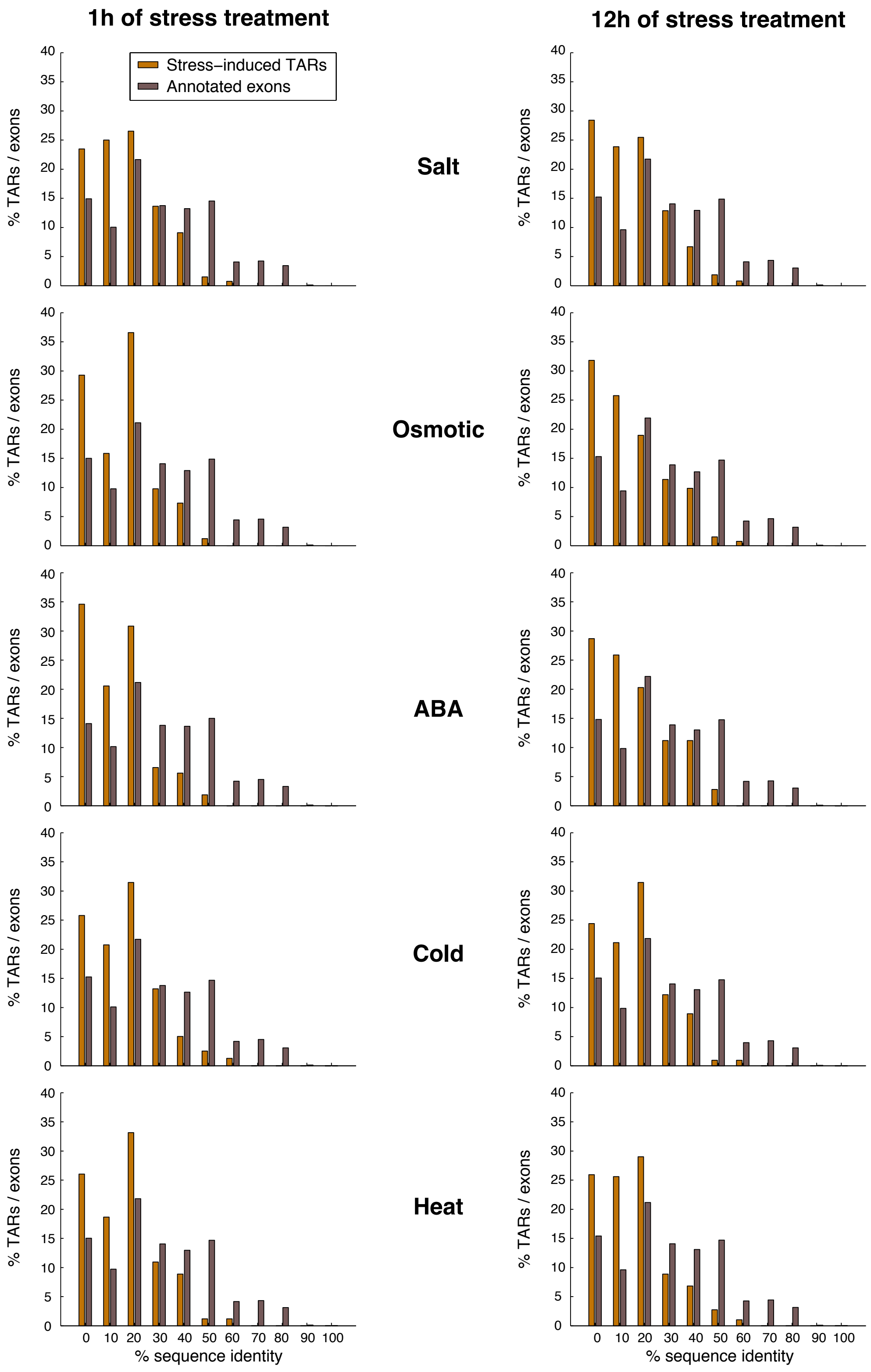


Figure 4.8: Histograms of (normalized) sequence identity of stress-induced TARs. For each treatment and time point a comparable histogram of randomly sampled annotated exons is shown as well (see inset and Section 2.6.7).

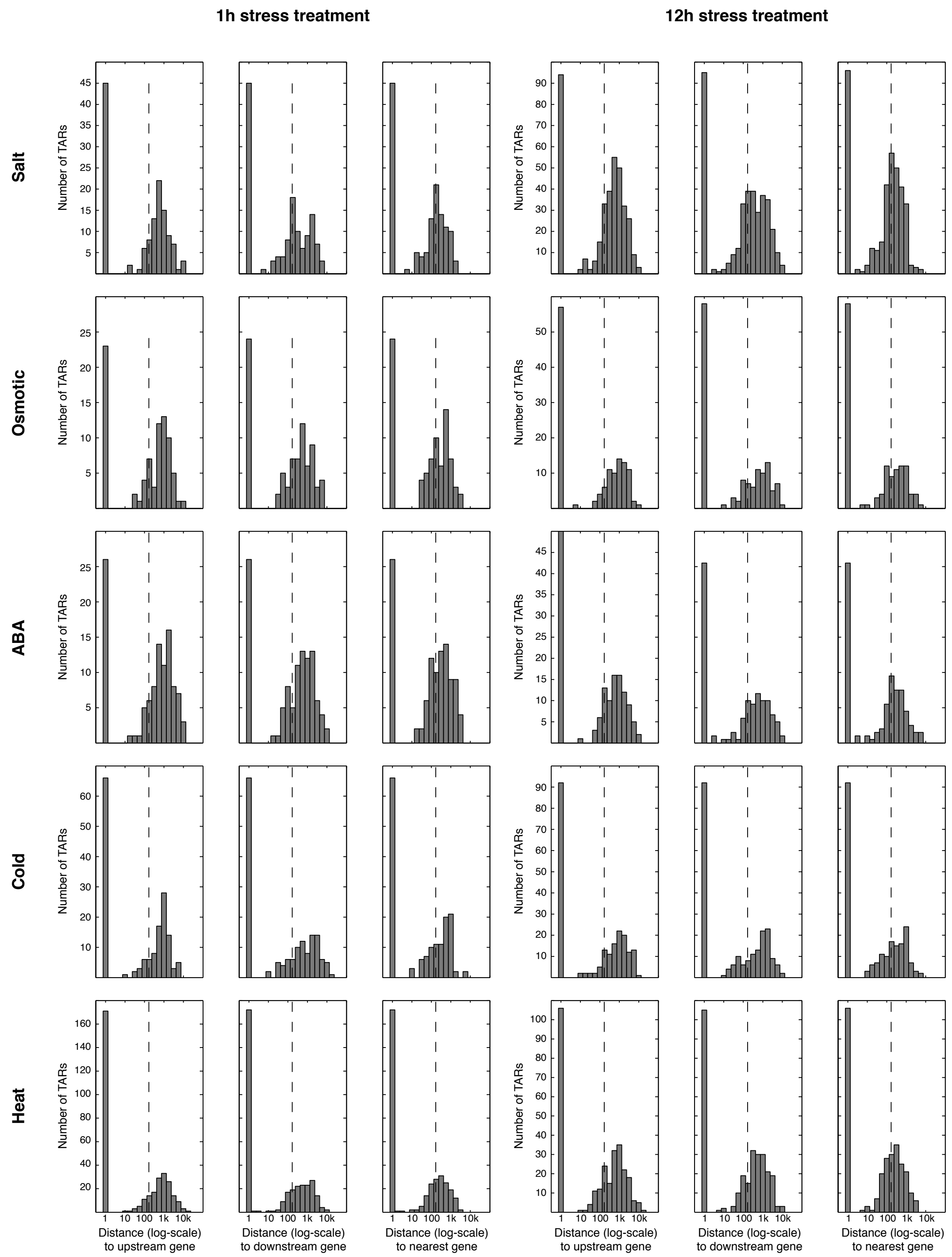


Figure 4.9: Histograms showing the genomic distance (log-scale) between uannotated, stress-induced TARs and nearest annotated genes by individual stress treatment and time point. A distance of 0 results from TARs overlapping with annotated exons or located in introns of annotated genes. For comparison, the median length of annotated *Arabidopsis* introns is indicated as vertical dashed line.

Publications

Peer-Reviewed Articles

Articles on which the core parts of this dissertation are based are marked with • and author contributions are indicated. Publications from which minor parts are taken or which are based on methods presented here are marked by ∘ and their relation to this work is outlined.

2007

∘ R. M. Clark, G. Schweikert,* C. Toomajian,* S. Ossowski,* **G. Zeller**,* P. Shinn, N. Warthmann, T. T. Hu, G. Fu, D. A. Hinds, H. Chen, K. A. Frazer, D. H. Huson, B. Schölkopf, M. Nordborg, G. Rätsch, J. R. Ecker, and D. Weigel. Common Sequence Polymorphisms Shaping Genetic Diversity in Arabidopsis thaliana. *Science*, 317:338–342, 2007.

Relation to this dissertation: In this work, the resequencing array data as well as a novel SNP calling algorithm (by G. Schweikert, G. Rätsch, and B. Schölkopf), its results and their biological implications were published. Additionally, the initial annotation of repetitive tiling probes was created for this project (by G. Zeller).

2008

• **G. Zeller**, S. R. Henz, S. Laubinger, D. Weigel, and G. Rätsch. Transcript Normalization and Segmentation of Tiling Array Data. *Pac. Symp. Biocomput.*, 13:527–538, 2008.

Author Contributions: G. Zeller and G. Rätsch conceptualized computational methods and developed normalization and segmentation algorithms; S. R. Henz evaluated competing methods; S. Laubinger generated tiling array data; S. Laubinger and D. Weigel helped with data interpretation; G. Zeller, S. R. Henz and G. Rätsch wrote the manuscript with contributions from S. Laubinger and D. Weigel.

• **G. Zeller**, R. M. Clark,* K. Schneeberger,* A. Bohlen, D. Weigel, and G. Rätsch. Detecting Polymorphic Regions in Arabidopsis thaliana with Resequencing Microarrays. *Genome Res.*, 18:918–929, 2008.

Author Contributions: G. Zeller and G. Rätsch developed the algorithm for detecting polymorphic regions (PRs); G. Zeller compared PRs to other polymorphism data with help from R. M. Clark; K. Schneeberger analyzed PR distribution; R. M. Clark, G. Zeller and K. Schneeberger and G. Rätsch interpreted data; A. Bohlen performed validation experiments; G. Rätsch, D. Weigel, R. M. Clark and G. Zeller conceptualized study; R. M. Clark, G. Zeller, G. Rätsch, K. Schneeberger and D. Weigel wrote manuscript.

*contributed equally

S. Laubinger, T. Sachsenberg, **G. Zeller**, W. Busch, J. U. Lohmann, G. Rätsch, and D. Weigel. Dual Roles of the Nuclear Cap-binding Complex and SERRATE in pre-mRNA Splicing and microRNA Processing in Arabidopsis thaliana. *Proc. Natl. Acad. Sci. U.S.A.*, 105:8795–8800, 2008.

• S. Laubinger, **G. Zeller**, S. R. Henz, T. Sachsenberg, C. K. Widmer, N. Naouar, M. Vuylsteke, B. Schölkopf, G. Rätsch, and D. Weigel. At-TAX: A Whole Genome Tiling Array Resource for Developmental Expression Analysis and Transcript Identification in Arabidopsis thaliana. *Genome Biol.*, 9:R112, 2008.
Author Contributions: S. Laubinger generated tiling array data, analyzed data and performed validation experiments; G. Zeller performed *de novo* transcript identification and analyzed results; G. Zeller generated probe annotations with help from N. Naouar, S. R. Henz and T. Sachsenberg; S. R. Henz performed expression analysis; T. Sachsenberg and C. K. Widmer developed visualization tools; D. Weigel, G. Rätsch, B. Schölkopf and M. Vuylsteke conceptualized study; S. Laubinger wrote manuscript with help from G. Zeller, D. Weigel, S. R. Henz and G. Rätsch.

○ N. Naouar, K. Vandepoele, T. Lammens, T. Casneuf, **G. Zeller**, P. van Hummelen, D. Weigel, G. Rätsch, D. Inzé, M. Kuiper, L. D. Veylder, and M. Vuylsteke. Quantitative RNA Expression Analysis with Affymetrix Tiling 1.0R Arrays Identifies New E2F Target Genes. *Plant J.*, 57:184–194, 2008.
Relation to this dissertation: In this work, the same type of tiling array data were analyzed using the custom tiling probe annotation created for the At-TAX project (by G. Zeller).

2009

S. Plantegenet, J. Weber, D. R. Goldstein, **G. Zeller**, C. Nussbaumer, J. Thomas, D. Weigel, K. Harshmann, and C. S. Hardtke. Comprehensive analysis of Arabidopsis Expression Level Polymorphisms with Simple Inheritance. *Mol. Syst. Biol.*, 5:242, 2009.

• **G. Zeller**,* S. Henz,* C. Widmer, T. Sachsenberg, G. Rätsch, D. Weigel, and S. Laubinger. Stress-induced Changes in the Arabidopsis thaliana Transcriptome Analyzed Using Whole Genome Tiling Arrays. *Plant J.*, 58:1068–1082, 2009.
Author Contributions: G. Zeller performed *de novo* transcript identification and analyzed transcript and expression data; S. R. Henz performed expression analysis; T. Sachsenberg and C. K. Widmer developed visualization tools; S. Laubinger generated tiling array data, analyzed data and performed validation experiments; S. Laubinger, D. Weigel and G. Rätsch conceptualized study; S. Laubinger wrote manuscript with help from G. Zeller, D. Weigel and S. R. Henz and G. Rätsch.

G. Schweikert, J. Behr, A. Zien, **G. Zeller**, J. Eichner, S. Sonnenburg, G. Rätsch. mGene.web: A Web Service for Accurate Computational Gene Finding *Nucl. Acids Res., Web Server Issue*, 37:W312–W316, 2009.

○ K. L. McNally, K. L. Childs, R. Bohnert, R. M. Davidson, K. Zhao, V. Ulat, **G. Zeller**, R. M. Clark, D. R. Hoen, T. E. Bureau, R. Stokowski, D. G. Ballinger, K. A. Frazer, D. R. Cox, B. Padhukasahasram, C. D. Bustamante, D. Weigel, D. J. Mackill, R. M. Bruskiewich, G. Rätsch, C. R. Buell, H. Leung, and J. E. Leach. Genome-wide SNP Variation Reveals Relationships Among Landraces and Modern Varieties of Rice. *Proc. Natl. Acad. Sci. U.S.A.*, 106:12273–12278, 2009.
Relation to this dissertation: For analyzing rice resequencing data similar to those generated for *Arabidopsis*, normalization, SNP calling, and PR detection methods were adopted and extended (by R. Bohnert with help from G. Zeller, G. Rätsch, and R. M. Clark).

G. Schweikert, A. Zien*, **G. Zeller***, J. Behr, C. Dieterich, C. S. Ong, P. Philips, F. De Bona, L. Hartmann, A. Bohlen, N. Krüger, S. Sonnenburg, and G. Rätsch. mGene: Accurate SVM-Based Gene Finding with Application to Nematode Genomes. *Advance access publication in Genome Res.*, 2009.

Unpublished

S. Laubinger*, **G. Zeller***, S. R. Henz, T. Sachsenberg, G. Rätsch, and D. Weigel. Global Effects of the Small RNA Biogenesis Machinery on the Arabidopsis Transcriptome. *Submitted to Proc. Natl. Acad. Sci. U.S.A.*, 2009.

Oral Presentations at International Meetings

2006

Machine Learning Algorithms for Polymorphism Detection, *presented (standing in for G. Schweikert) at the 2nd Student Council Symposium (SCS) jointly held with the International Conference on Intelligent Systems for Molecular Biology (ISMB),* 2006.

2008

Transcript Normalization and Segmentation of Tiling Array Data, *presented at the Pacific Symposium on Biocomputing,* 2008.

Re-Annotating the *Arabidopsis* Transcriptome with Tiling Arrays, *presented at the 4th Student Council Symposium (SCS) jointly held with the International Conference on Intelligent Systems for Molecular Biology (ISMB),* 2008.

2009

Characterizing Transcriptome Plasticity Using Whole-genome Tiling Arrays and Machine Learning, *Highlight presentation at the International Conference on Intelligent Systems for Molecular Biology (ISMB) & European Conference on Computational Biology (ECCB),* 2009.