Polymorphism Detection in Rice

Diplomarbeit im Studiengang Bioinformatik

vorgelegt von

Regina Bohnert

Erstgutachter: Prof. Dr. Daniel Huson[1,2]
Zweitgutachter: Dr. Gunnar Rätsch[3]

Wilhelm-Schickard-Institut für Informatik
Fakultät für Informations- und Kognitionswissenschaften
Universität Tübingen

Tübingen, im Oktober 2007

[1]Wilhelm-Schickard-Institut für Informatik
[2]Zentrum für Bioinformatik Tübingen
[3]Friedrich-Miescher-Labor der Max-Planck-Gesellschaft, Tübingen

Polymorphism Detection in Rice

Regina Bohnert

Abstract

The major breakthrough at the turn of the millennium was the completion of genome sequences for many species, including humans, worms and rice. Since then, it has been realized that making use of these resources requires as much knowledge about sequences shared between members of the same species, that is, the information about sequences that distinguish individuals with different traits.

Recently, the International Rice Functional Genomics Consortium began a project to investigate sequence variation within and between rice strains on a genome-wide scale. The central goals of this project is the resequencing of the non-repetitive portion of the genomes of 20 diverse strains using high-density oligonucleotide microarrays. The resulting knowledge about sequence variation will be the basis for further evolutionary and functional studies of rice. A long-term aim is the identification of sequence variants that confer advantageous traits and that are thus targets for crop improvement through breeding.

This Diplom thesis contributed to the rice resequencing project by analysing the rich set of hybridisation data from the resequencing arrays in order to reveal hundreds of thousands of single nucleotide polymorphisms (SNPs). The application of the raw data for further computational analyses required several preprocessing steps. In a first step, we applied quantile normalisation to enable the comparison of hybridisation intensities across experiments. In a second step, we addressed the problem of cross-hybridisation by identifying repetitive 25-mers in the reference rice genome. The results of the latter analysis confirm previous reports that single nucleotide substitutions, as well as one base pair insertions/deletions (indels), have a major influence on the performance of this method.

An important part of this thesis is the compilation of a set of reference polymorphisms, which were identified by dideoxy sequencing, and to assess the quality of these reference data, which are required for subsequent SNP discovery and its evaluation. More than 3,500 fragments of average length of 557 bp had been dideoxy-sequenced from up to three strains, resulting in set of 1,755,395 positions across all strains. Polymorphic sites were be detected at 9,499 positions with 6,510 non-redundant SNP positions.

In the last step, a technique based on Support Vector Machines was applied to identify SNPs from the hybridisation data. In comparison to a model based (MB) SNP calling approach implemented by Perlegen Sciences, the machine learning (ML) method was superior by recovering 20.9% of all known SNPs at a false discovery rate (FDR) of 8.3%, compared to 13.9% and 7.8%, respectively, for the MB approach. Both methods together predicted 1,930,400 SNPs at 421,965 positions. These predictions can now be used for subsequent analyses; for example, to identify potential functional effects of the SNPs, to discover regions of the genome shared between otherwise distant strains, reflecting certain historical breeding events, or to describe genome-wide haplotype structure (i.e. construct a rice HapMap).

Acknowledgement

First, I want to thank Daniel Huson, Gunnar Rätsch and Detlef Weigel for giving me the opportunity to contribute to this excellent research project and for supervision of this Diplom thesis.

Many thanks to Georg Zeller for his support, advice and valuable discussions throughout the project. Moreover, I am very grateful to him for reading and reviewing the manuscript of this thesis. I would also like to thank Gabriele Schweikert for providing code and support on the SNP calling problem.

Furthermore, thanks to Richard Clark at the Max Planck Institute for Developmental Biology in Tübingen, Kevin Childs and Robin Buell at The Institute for Genomic Research in the USA for valuable collaboration.

Thanks also go to all my colleagues in the Rätsch group who have helped to create a positive working environment.

Contents

Abstract iii

Acknowledgement v

1 Introduction 1
- 1.1 Rice – an Important Food Source and Model Organism . . . 1
- 1.2 Sequence Variations . . . 2
- 1.3 Variation Studies and their Benefits . . . 3
- 1.4 The Rice Resequencing Project . . . 5
- 1.5 Aims of the Thesis . . . 7
- 1.6 Thesis Structure . . . 7

2 The Resequencing Data 9
- 2.1 Tiling Arrays . . . 9
- 2.2 Array Design . . . 11
- 2.3 Processing the Raw Data . . . 12
- 2.4 Quality Assessment . . . 15

3 Annotation of Repetitive k-mers in the Reference Genome 17
- 3.1 Motivation . . . 17
- 3.2 Definitions . . . 19
- 3.3 Approach . . . 21
- 3.4 Results and Discussion . . . 25

4 The Dideoxy Sequencing Data 31
- 4.1 Sample Selection . . . 31
- 4.2 Processing the Raw Data . . . 33
- 4.3 Quality Assessment of MB SNP calls . . . 34
- 4.4 Quality Assessment of the Dideoxy Sequencing Data . . . 37

5 SNP Calling 43
- 5.1 A Two-Layered Machine Learning Method for SNP Calling . . . 43
- 5.2 Results . . . 51
- 5.3 Conclusions . . . 55

6 Conclusion and Future Work 59

References 61

Appendix 65

1 Introduction

1.1 Rice – an Important Food Source and Model Organism

For some thousands of years rice has been used by human as one of the main food sources. Derived from archaeological rice remains dating to 8,000 BC it is believed that domestification and cultivation of the crop have been taken place since 10,000 years now (Sweeney and McCouch, 2007). Today, rice is the most important source for human caloric intake, making up 20% of the calorie supply (International Rice Research Institute, 2007a) and feeding millions of people daily.

Due to its relevance as calorie provider and other important properties such as having a reasonable genome size, being closely related to other cereal species and representing the plant family of grasses, rice has been heavily studied in recent years. It became the second most prominent model organism of plants besides *Arabidopsis*, providing various insights on monocotyledons,[1] one of the two major groups of flowering plants. The more detailed understanding and findings on the molecular assembly of phenotypic rice varieties will be essential for future improvement in rice cultivation and breeding.

Rice is scientifically classified as the genus *Oryza*, comprising both wild and domesticated rices. There are two distinct types of domesticated rice: *Oryza sativa* which is found all over the world with a high accumulation in Asia and *Oryza glaberrima*, a rice species growing in West Africa (Sweeney and McCouch, 2007). By previous classification, Asian rice was splitted up into the two subspecies *indica* and *japonica* of which the crosses are infertile. According to morphological as well as recent genetic studies by e.g. Garris et al. (2005), Asian rice can be subdivided into five major populations: *aus* and *indica*, defining the traditional *indica* subspecies, and *temperate japonica*, *tropical japonica* and *aromatic* forming the *japonica* subspecies. The estimation of genetic divergence of the two main subspecies *indica* and *japonica* to 100,000 years suggests that two or more independent domestication events in *Oryza sativa* happened to form today's subpopulations (Sweeney and McCouch, 2007).

In 2002, two sequence drafts of the rice genome were published. The first release by Goff et al. (2002) is based on a whole genome assembly (WGA) with 6 × coverage of the strain Nipponbare, belonging to the *temperate japonica* subpopulation. Secondly, Yu et al. (2002) sequenced a member of the *indica* subpopulation, 93-11, having a 6.38 × WGA. Three years after that, in 2005, a finished genome sequence of Nipponbare was published by the International Rice Genome Sequencing Project (IRGSP), with a tenfold sequence coverage and with an estimated quality of less than one error per 10,000 bases (International Rice Genome Sequencing Project, 2005). Analyses of the IRGSP genome version resulted in the estimation of the total rice genome size of 389 Mega bases (Mb), distributed over 12 chromosomes with a GC content of 43% (The Rice Annotation Project, 2007), of which 370 Mb (95%) were

[1] Monocotyledons are the group of flowering plants having one embryonic leaf, whereas dicotyledons, to which e.g. *Arabidopsis* belongs, develop seeds with typically two embryonic leaves.

covered by the assembly of the sequenced contigs. Recently, efforts were undertaken to annotate the IRGSP rice genome by the Rice Annotation Project (The Rice Annotation Project, 2007). Using cDNAs and expressed sequence tags (ESTs) the number of genes were conservatively estimated to 32,000. In contrast to that, approximately 56,300 genes are expected based on an annotation using both transcript assemblies and computational predictions (see The Institute for Genome Research, 2007; Rensink and Buell, 2004).

1.2 Sequence Variations

With the genome sequence of two rice subspecies which diverged long time ago, first studies were made on a molecular level investigating the sites which are distinct between the *O. sativa indica* and *japonica* subspecies (International Rice Genome Sequencing Project, 2005). Differences in the genome sequence within a species are referred by the term sequence variations. New sequence variations occur when the sequence composition and arrangement is altered by mutation events on gene, chromosomal and genomic level. Such mutations can lead to phenotypic changes, diseases or have even lethal consequences.

Single nucleotide polymorphisms (SNPs) are the most abundant class of sequence variations. They are caused by mutations which alter one single nucleotide in the genome of individuals of one species or in a chromosome pair of one individual (see Lodish et al., 2004, p. 396/7,



Figure 1.1: Rice is a plant of 1-1.8 m with long slender leaves. It proliferates by grain which are about 1 cm long and a few millimetres thick. The plant is native to Asian and African areas where the tropical and subtropical climate promotes its growth. The picture on the right hand side illustrates the variety of rice grains. The pictures are taken from the Wikipedia (2007) and the International Rice Research Institute (2007b) websites.

Chapter 9). A SNP can be located anywhere in the genome, but different effects can be observed depending on the sequence context. When a SNP is observed within a protein coding gene, the alternate nucleotide often causes a change in the protein sequence and can thus be relevant for phenotypic modifications. Changes in signal sites such as transcription start sites and splice sites can lead to alternate transcription and splicing behaviour. SNPs within intergenic regions and untranslated regions (UTRs) can have effects on the regulation of gene expression and on sequences encoding non-coding RNAs.

Length polymorphisms can be introduced by a number of repetitions of a few bases (see Lodish et al., 2004, p. 397, Chapter 9). Such polymorphisms are called simple sequence repeats (SSRs) or microsatellites. They originate from an event during DNA replication in which a part of either the template or the newly synthesised strand is slipped.

Another type of longer changes in nucleotide sequences are referred as deletion and insertion polymorphisms, abbreviated as indels. A deletion is the loss of a sequence region which can have a length of a single nucleotide, a few nucleotides or large parts of the chromosome. Insertions are the opposite of deletions, describing the gain of sequence regions.

1.3 Variation Studies and their Benefits

Research on polymorphic patterns and sequence divergence between rice strains has been undertaken since the first drafts of the rice genome were published. The paper by Han and Xue (2003) summarises work on the comparison of selected homologous regions, covering less than 1% of the genome, between the *indica* strain Guangluai and the *japonica* strain Nipponbare. In another study by Liu et al. (2007), sequence variations in expressed sequence regions of the *indica* and *japonica* subspecies were investigated based on a collection of roughly 10,000 *indica* full-length cDNAs.

Moreover, there are several studies on genome-wide patterns of polymorphism. A polymorphism database based on the comparison of the *indica* strain 93-11 and the *japonica* strain Nipponbare was built by Shen et al. (2004) containing 1,703,176 SNPs and 479,406 indels with a SNP frequency of 3.5 SNPs/kb. Monna et al. (2006) investigated genomic fragments from predicted intergenic regions in eight *O. sativa* strains and a wild strain aiming to detect SNPs as well as indels. To explain sequence variation in domesticated rice, 111 randomly chosen gene fragments were sequenced for 72 domesticated *O. sativa* strains and 21 wild *O. rufipogon* strains by Caicedo et al. (2007). The authors concluded that either positive selection or a severe population bottleneck has strongly influenced the evolution of today's polymorphism patterns.

As all previous studies were limited to certain parts of the rice genome or to a small selection of strains, a whole-genome analysis on polymorphisms in diverse strains has been in great demand. Such a whole-genome collection of polymorphisms will further be essential for addressing questions on rice evolution and population genetics. Studies in these fields often investigate which SNPs are causal for the variation in complex traits. For this purpose, a polymorphism collection serves as a set of markers to detect regions of the genome in association with the trait of interest. Statistical methods such as quantitive trait locus (QTL) mapping approaches this question (e.g. Weigel and Nordborg, 2005). Another approach called linkage disequilibrium (LD) or association mapping aims to analyse the non-random association of alleles at more than two loci using the frequencies of defined allele combinations

(Syvänen, 2001). Influences of sequence variations on gene expression due to changes in the regulatory sequences and alternate amino acids in proteins can also be studied, crucial for the better understanding on the relation between genetic composition and phenotypic observations.

Once a collection of polymorphisms has been determined, a set of informative genotyping markers can be selected. For instance, a tool provided by Warthmann et al. (2007) can be applied for this purpose which facilitates SNP marker detection by extracting information on polymorphisms from multiple sequence alignments. For genotyping based on the usage of SNP markers, a number of biological assays have been established. The simplest and oldest SNP-genotyping technology, the restriction fragment length polymorphism (RFLP) method, makes use of different restriction enzymes with high affinity to distinct restrictions sites. At these consensus sequences, the enzymes cut the genome into fragments of which the lengths are determined by a gel assay. Only SNPs occurring at the restriction sites are detectable by this method as sites with altered consensus sequence are not recognised by the restriction enzyme (see Lodish et al., 2004, p. 397, Chapter 9). Other enzyme-based methods apply allele-specific polymerase chain reactions (PCRs) for which SNP-specific primers are designed to amplify a sequence at a distinct polymorphic site. In minisequencing single nucleotide primer extension, a primer is located immediately adjacent to a SNP. The primer is extended with a single nucleotide complementary to the SNP which can then be detected. Oligonucleotide ligation assays exploit the property that the enzyme ligase only connects perfectly matched probes. Additionally, sequencing methods such as Sanger dideoxy sequencing or Pyrosequencing are commonly used, but limited to sequence only small parts of the genome in a short time frame. Hybridisation methods, among them the widely used technology of DNA microarrays, rely on fluorescence detection of the DNA duplex formation between the sequence including a SNP and its complement.

As mentioned before, such SNP-genotyping technologies require a collection of known polymorphisms. An approach towards whole-genome SNP identification is the method of array-based resequencing. Here, the genomes of multiple strains of a certain species are resequenced with high-density oligonucleotide microarrays which are designed using a high-quality reference genome as a template. Comparing a given target strain to the reference sequence allows to detect differences between these two DNA sequences in a more efficient and cost-effective way than *de novo* sequencing using the Sanger method.

Such resequencing methods have been applied to discover SNPs in several organisms. Resequencing genomes of 24 human individuals from Europe, Africa and Asia and genotyping roughly 1.5 million SNPs, Hinds et al. (2005) described whole-genome patterns of common human DNA variations. Recently, resequencing projects for two other organisms were finished. Clark et al. (2007) identified and investigated common sequence polymorphisms of 20 diverse *Arabidopis thaliana* strains. This study resulted in the discovery of 1 million nonredundant SNPs. Similar methods were used for the mouse resequencing project in which 8.27 million SNPs densely distributed across the mouse genome were identified for 15 mouse strains (Frazer et al., 2007).

1.4 The Rice Resequencing Project

To investigate genetic variation within and between diverse rice strains, the International Rice Functional Genomics Consortium (IRFGC) set up a project wherein several research groups, among others groups from The Institute for Genomic Research (TIGR) in the USA, the International Rice Research Institute (IRRI) on the Philippines and the Max Planck Institute for Developmental Biology in Tübingen, Germany, have been involved (McNally et al., 2006). For resequencing with high-density oligonucleotide microarrays 20 different rice strains were selected. Criteria for selection of the strains were the strain's value in breeding and genetic studies and the relative diversity to each other (McNally et al., 2006). Figure 1.2 gives an idea on the evolutionary distance and geographic representation of the 20 rice strains which are described in more detail in Table 1.1.

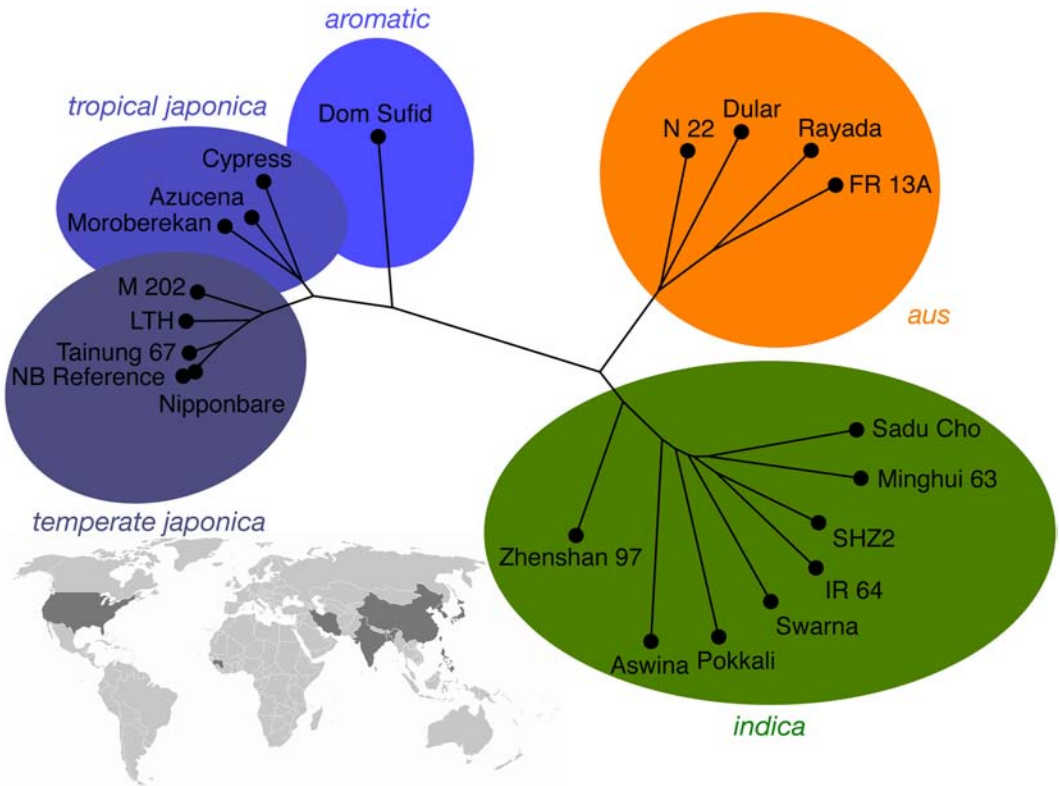


Figure 1.2: The evolutionary distance of the 20 selected rice strains is illustrated. Blue colouring denotes strains of the *japonica* subspecies. The *aus* strains are coloured in orange and all strains belonging to the *indica* subspecies are shown in green. The world map at the bottom illustrates the geographic representation of the 20 different strains, coloured in dark grey. Note that most of the strains originate from areas in Asia.

Table 1.1: The 20 rice strains selected for resequencing.

Strain	Origin	Species	Agronomic Attributes
Aswina	Bangladesh	*indica*	Deep-water rice.
Azucena	The Philippines	*trop. japonica*	Grain-iron quantitative trait loci. Deep root distribution.
Cypress	USA	*trop. japonica*	Good grain quality. Cold tolerant.
Dom Sufid	Iran	*aromatic*	Basmati plant type. Aromatic rice.
Dular	India	*aus*	Possible β-carotene donor. Red pericarp.
FR 13A	India	*aus*	Submergence tolerance. Red pericarp.
IR 64	The Philippines	*indica*	Multiple stress tolerance (diseases and insects). Progenitor of many breeding and mapping populations.
Li-Jiang-Xin-Tuan-Hei-Gu (LTH)	China	*temp. japonica*	High disease susceptibility. Cold tolerant.
M 202	USA	*temp. japonica*	Popular variety.
Minghui 63	China	*indica*	
Moroberekan	Guinea	*trop. japonica*	Multiple advanced breeding populations. Drought tolerant.
N 22	India	*aus*	Iron, red pericarp. Heat tolerant and considered drought tolerant.
Nipponbare	Japan	*temp. japonica*	Served as reference strain.
Pokkali	India	*indica*	Salt tolerant.
Rayada	Bangladesh	*aus*	Deep-water rice.
Sadu Cho	Korea	*indica*	Long grain and indica-type endosperm.
Shan-Huang Zhan-2 (SHZ2)	China	*indica*	Disease-resistant, high-yielding variety. In the pedigrees of many varieties in south China.
Swarna	India	*indica*	High yield potential, wide adaptability. Widely planted variety in India.
Tainung 67	Taiwan	*temp. japonica*	
Zhenshan 97	China	*indica*	

1.5 Aims of the Thesis

Analyses of the hybridisation data generated by the resequencing of the 20 rice strains require efficient and accurate computational methods. Approaches similar to the *Arabidopsis* resequencing project were realised for SNP discovery in rice.

To contribute to the rice resequencing project in this regard, this thesis aims towards the following goals:

- Improvement and extension of the method for annotation of repetitive 25-mers in the reference genome Nipponbare
- Design and quality assessment of a polymorphism set which was derived by dideoxy sequencing
- Adaptation and extension of the SNP calling algorithm applied in the *Arabidopsis* resequencing project (Clark et al., 2007) to rice

1.6 Thesis Structure

Having introduced the biological problem in this chapter, we focus on the resequencing data generated by tiling microarrays in the next chapter. After giving an overview over the general design and manufacturing of microarrays and tiling arrays in particular, we present the experimental design for the rice resequencing project. Furthermore, processing and quality control of the hybridisation data are described.

In Chapter 3, we address the question how we can efficiently detect repetitive k-mers in the rice reference genome. We begin with motivating the demand of such a k-mer analysis and then go into details on the implementation of the k-mer analyser. At the end of the chapter, we present results and the application of the k-mer data.

Subsequently in Chapter 4, we describe the dideoxy sequencing data which was utilised for training the machine learning algorithms and for quality assessment of polymorphism predictions. For this purpose, we illustrate the selection procedure of the samples for sequencing and further processing of the sequencing data. The bigger part of this chapter focuses on how we assessed the quality of the dideoxy sequencing data and the consequences for further application of the data.

Finally in Chapter 5, we concentrate on addressing the question how we can exploit the resequencing data for identification of SNPs across all non-reference strains. We therefore first introduce the concept of Support Vector Machines (SVMs) and then present our two-layered SNP calling approach based on SVMs. Thereafter, results and performance evaluation in comparison to another SNP calling method are given.

2 The Resequencing Data

In this chapter, we describe the generation, processing and quality control of the resequencing data. We first give an overview over the characteristics of tiling arrays and their manufacturing. Thereafter, details on the design of the resequencing arrays and experiments are illustrated. We then describe how we processed the raw hybridisation data obtained by resequencing experiments in order to use the data in further analyses. The quality assessment of the hybridisation data is described in the end of this chapter.

2.1 Tiling Arrays

DNA microarrays are a well-established technology and have been in use for over ten years, permanently in further development, improvement and optimisation. They allow to determine complementary binding of sample DNA by DNA molecules immobilised to a solid surface such as glass (Ramsay, 1998).

Microarray technologies can be distinguished by probe type, i.e. whether longer nucleotide molecules, such as cDNA or PCR products, or oligonucleotides are used as probes. Another discrimination is by the manufacture type: The probes are either spotted and immobilised onto the array surface by robots or an array of oligonucleotides is synthesised *in-situ* (see Lodish et al., 2004, p. 385/6 , Chapter 9). A method for the manufacture of *in-situ* arrays is shown in Figure 2.1.

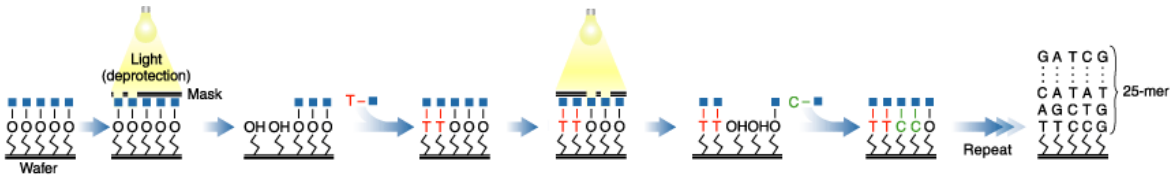


Figure 2.1: The process of array manufacturing using photolithography. A light-sensitive chemical compound is coated on the array surface, preventing the attachment of the first nucleotide onto the array. With the usage of lithographic masks light is permitted to hit onto specific locations on the surface. In an iterative process, a solution of nucleotides containing either the base Adenine (`A`), Cytosine (`C`), Guanine (`G`) or Thymine (`T`) is added in order to attach to those locations which were deprotected by the light. The coupled nucleotide has also a light-sensitive protecting group. Thus, the cycle can be repeated until the desired oligonucleotide length (25 to 100 nt) is reached. The picture is taken from Affymetrix (2007).

Depending on the type of the genetic study, the targets interrogated with DNA microarrays are different. Microarrays have been widely applied to investigate gene expression under different conditions. Here, cellular RNA is used for the identification of genes which are differentially expressed. For the study of sequence variation in one particular population, genomic DNA from different individuals of this population serves as a target template. The RNA or DNA sample is purified to obtain single-stranded DNA which is labelled with a fluorescent compound and added on the array for hybridisation. Afterwards, the array is washed in order to remove unspecifically bound sequences. By scanning with a fluorescence microscope, the fluorescence intensities for each probe-target pair are measured reflecting the degree of hybridisation of a certain probe to DNA from the sample.

The availability of entirely sequenced genomes and the miniaturisation of the arrays have enabled the application of microarrays for whole-genome investigations. Such high-density whole-genome oligonucleotide arrays are useful for a bulk of genetic studies, among them polymorphism analyses via genome resequencing (Mockler et al., 2005). As denoted in their naming, these arrays have a high density of a few millions of so-called unique features, each feature comprising in turn a large number of copies of a distinct probe sequence.

For resequencing purposes, a reference genome serves as a template for designing a set of tiled, overlapping oligonucleotide probes. There are eight probes, four representing the forward and reverse strand, respectively, for each nucleotide of the genome of interest in total. One quartet includes the reference probe and three other probes where the central base in the probe is altered to one of the three other bases. Hence, the number of probes is an eight fold of the tiled sequence length. Ideally, a SNP at the centre position of a probe causes a reduced hybridisation signal for the reference probe, but an increased signal for the mismatch probe and can be detected in that way. Figure 2.2 illustrates the tiling design for SNP calling.

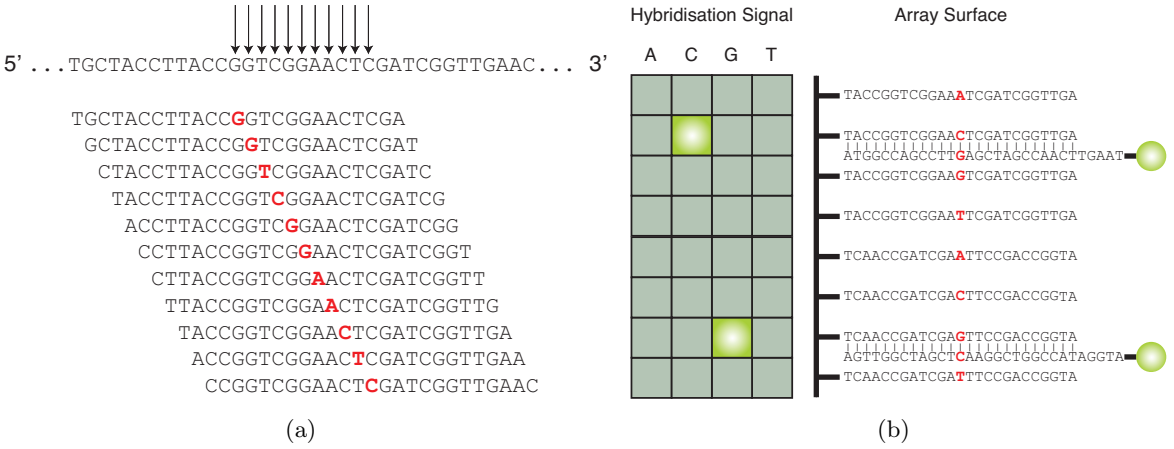


Figure 2.2: The design of high-density tiling arrays for resequencing. (a) Each position in the genome is represented by a group of 25-mers which comprises the sequence from 12 nucleotides upstream and 12 bases downstream of the position of interest. (b) Resequencing arrays are designed in a way that eight probes are synthesised on the array for each position of interest. One quartet contains four probes of one strand with each of the four bases at the centre position of the probe.

Although whole-genome tiling arrays provide the benefit of having signal information for the entire genome, there are some problematic issues to address. Different hybridisation affinities mainly due to probe sequence effects and unbalanced concentrations of the target sequences give rise to analysis problems. As the probe itself does not contain any information on its context, i.e. where it is located in the genome, repetitive oligonucleotides can distort the signal, a phenomenon termed by cross-hybridisation. Methods approaching some of these problems are described in Section 2.3 of this Chapter and in Chapter 3. Another drawback of resequencing with microarrays is the loss of sites at which the signal information is either too noisy or the signal strength is too low to reliably determine the base. The quality of the hybridisation data generated for the rice resequencing project is assessed in Section 2.4 at the end of this chapter.

2.2 Array Design

At a first step, the reference sequence (*O. sativa* ssp. *japonica* Nipponbare, IRGSP build 4 (International Rice Genome Sequencing Project, 2005)) was used in order to build a repeat-masked version of the rice genome. Highly repetitive sequence parts were excluded as they very likely lead to useless hybridisation signals due to cross-hybridisation events. From this genome version 11,343 non-overlapping sequence fragments were selected, covering 117,834,417 bases in total of which 100,104,806 Mb were non-repetitive (Perlegen Sciences, 2007). The non-repetitive parts served as a template for generating the arrays probes using the tiling strategy with 1 base pair (bp) resolution. Thus, probes representing roughly 32% of the rice genome were synthesised on the arrays. Due to the design of eight distinct features per interrogated site there were 800,838,448 features all together.

The probes were distributed over six *wafers*. One wafer contained only one array serving as a replicated development array to test experimental conditions for sample preparation, hybridisation, staining and scanning of the arrays. All of the other five wafers were designed in the same way, each containing 49 distinct arrays. Hence, the total number of arrays manufactured for the rice project was 246 for each strain.

For the generation of the target DNA 13,582 amplicons were amplified for each of the 20 strains by long range polymerase chain reaction (LR-PCR), having 271,640 PCR reactions in total. The amplified sequences were split up into 16 pools for each strain. The number of amplicons in each pool varied from 908 to 1,039, with the exception of the pool for the developmental array with 76 amplicons. Figure 2.3 shows the pooling for a 49-array wafer. The mean amplicon length was 1,764 bp with an standard deviation of 1,922 bp. 50% of all amplicons were shorter than 1,000 bp, with an accumulation of amplicons in the range of 100 to 200 bp.

01	02	03	04	05	06	07
08	09	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31	32	33	34	35
36	37	38	39	40	41	42
43	44	45	46	47	48	49

Figure 2.3: The picture illustrates how one wafer was divided into three sections, each corresponding to one pool. Amplicons from the first pool were hybridised to the arrays 1 to 16 (coloured in blue), the second amplicon pool to the arrays 17 to 32 (green) and the third amplicon pool to the remaining 17 arrays (orange).

2.3 Processing the Raw Data

2.3.1 Parsing Relevant Information from the Trace Files

The hybridisation data was converted into an compressed trace file format named ZTR by Perlegen Sciences using the Staden trace file IO library.[1] This trace file format was developed to produce compact files from DNA trace data suitable for large-scale storage and efficient transfer (Bonfield and Staden, 2002). A trace file contained information on the raw base calls, the base confidence values, the intensity measurements for each of the four bases, the offset of the base calls relative to the trace coordinates and several automatically generated and specific comments on e.g. run date and experiment number.

We used the Staden package to decompress the ZTR-files into plain text files. Each trace file contained data for one contiguous fragment of tiled sequence for one strand. These text files were parsed in order to collect and extract hybridisation intensities and quality scores on both strands for each site tiled on the arrays. The hybridisation intensities were mean fluorescence measurements, ranging from 0 to 4,095. Furthermore, a raw base call referred to the base for which the hybridisation intensity is highest within a probe quartet. Data for the reverse strand was complemented, such that each base call was in terms of the forward strand. Additionally, Perlegen Sciences provided information on the estimated error probability e of calling a certain base. These error probabilities are comparable to those used as a quality measure in dideoxy sequencing. They were computed using an algorithm similar to Phred (see Ewing and Green, 1998) considering relative intensities of the probes with highest and second highest intensity values in respect to surrounding base calls. These error probabilities were converted into quality scores by a log-transformation

$$q = -10 \cdot \log_{10}(e).$$

[1] `http://staden.sourceforge.net`

Thus, a high quality scores reflects a low error probability. Due to the characteristics of the scoring algorithm, the range of quality score values was limited to a set of discrete scores, ranging from 0 to 42 and skipping the values 1, 33, 40 and 41.

After parsing the ZTR-files the information on the hybridisation data was stored into 20 tables in a MySQL database. A key was added identifying each table row with a unique id allowing efficient sorting and joins across the tables.

2.3.2 Quantile Normalisation

Normalisation of hybridisation data is an important preprocessing step to obtain comparability of the data generated by multiple array experiments. This is motivated by the fact that systematic trends cause technical variation which has to be removed for further data interpretation to detect biological variation. Technical variation can arise due to differences in sample preparation, e.g. the usage of different dyes for labelling, and in array experimental procedures such as variability during hybridisation or variability introduced by differing experimentators. Scanner differences are also problematic in processing microarray data.

A simple approach to remove artificial variation is to scale the intensities such that each array has the same average value. Normalisation can also be achieved by linear regression when two hybridisation experiments are compared. The drawback of both methods is the assumption of a linear relationship between arrays. An approach based on non-linear relationship is realised with locally weighted (lowess) regression. In the paper by Schadt et al. (2001) an algorithm was proposed using invariant difference selection to fit a non-linear relationsship between a baseline and other arrays.

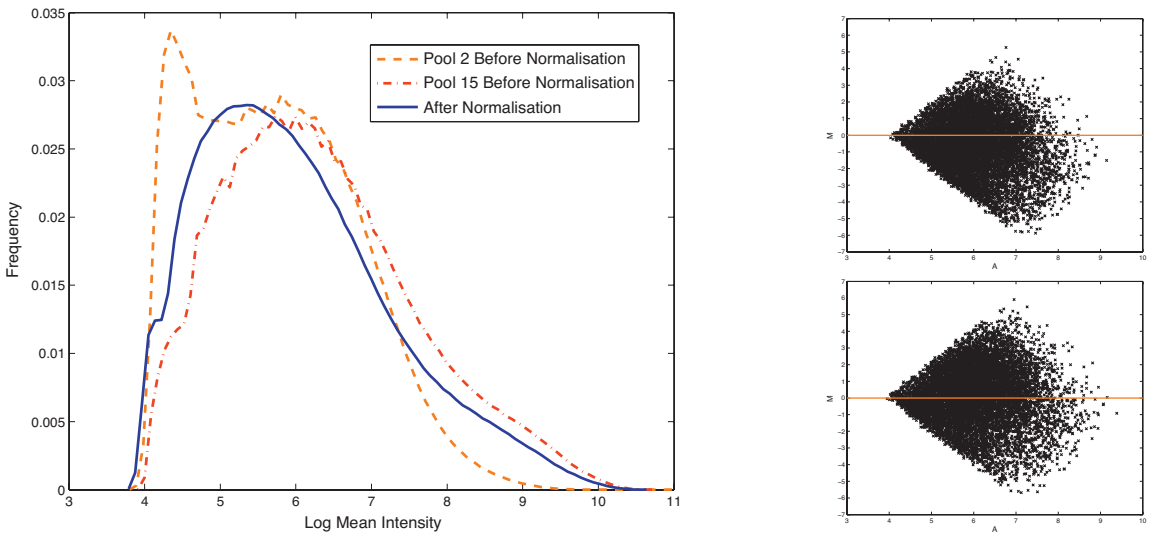
To normalise the rice resequencing data we decided to apply a method which does not depend on the choice of a baseline array, but is based on quantiles. The so-called quantile normalisation was introduced by Bolstad et al. (2003) and assumes that the intensity distribution for each array should be identical. Thus, the method aims towards adjusting the probe intensities in a way that their distribution is the same for each array in a set of arrays. The motivation of this approach arose from quantile-quantile (QQ) plots in which the quantiles of examples underlying a certain distribution are plotted against the quantiles of a second set of examples. If the plot shows a straight diagonal line, the same distribution underlies both example sets. In quantile normalisation the idea of QQ plots is generalised from two distribution to n distributions such that if all n sets of examples are distributed in the same way, then the n-dimensional plot of their quantiles gives a straight line along the unit vector $\mathbf{d} = (\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}})$. Transfered to the problem of array normalisation the question is to find a mapping ϕ which projects the points in the quantile plot onto this straight line. More formally, the projection of $\mathbf{q}$ onto $\mathbf{d}$ is defined as

$$\phi_{\mathbf{d}\ \mathbf{q}_k} = \left(\frac{1}{n} \sum_{j=1}^{n} q_{kj}, \ldots, \frac{1}{n} \sum_{j=1}^{n} q_{kj} \right)$$

where $\mathbf{q}_k = (q_{k1}, \ldots, q_{kn}), k = 1, \ldots, p$ is the vector of the kth quantiles for all n arrays, p denotes the total number of probes on each array.

The algorithm realising quantile normalisation consists of several steps. The data vectors of each array having length p are first pooled into a matrix X. The dimension of X is $p \times n$, i.e. each array is represented by one column. The intensity set of each column is sorted by a permutation Π_i, $i = 1, \ldots n$ resulting in a sorted matrix X'. Each entry in X' is then substituted by the mean across the rows of X'. Rearranging each column by Π_i^{-1}, X' is finally transformed to the matrix of normalised intensities.

The quantile normalisation algorithm was implemented and applied by Georg Zeller to normalise the rice resequencing data on the unit of amplicon pools across all strains. As the amplicon pools do not all contain the same number of probes, each pool was filled up with fake intensities to have a global identical pool size. Confounding effects were minimised by sampling the fake intensities from the observed distributions. Pool 0 on wafer 0 (developmental array) was excluded from normalisation since the number of probes in that pool was 13-fold lower than in the other 15 pools. Therefore, the sites queried by the development array were omitted in further analysis and the total number of sites was reduced to 99,725,727. The effect of normalisation for a small region is visualised in Figure 2.4.



(a) (b)

Figure 2.4: The effect of normalisation on the hybridisation intensity distribution. (a) The figure shows the different intensity distributions of pool 2 and 15 in dashed lines. The distribution after normalisation is shown in blue. (b) The MA plot visualises average intensities of two distributions h_1 and h_2 $A = 0.5 \cdot (\log h_1 + \log h_2)$ versus the difference of these intensities $M = \log h_1 - \log h_2$. The more the data points are clustered around the horizontal line through the origin, the more similar are the distributions. Before normalisation (figure at the top) the distributions are different indicated by the skew to lower M values which is corrected after normalisation (figure at the bottom).

2.4 Quality Assessment

In order to have a picture of the hybridisation data quality, we performed simple statistics and visualised them in various figures. We first examined how well the bases were called on both strands for the reference strain Nipponbare. We therefore counted the number of positions p for which the base call on the forward $B_+(p)$ and reverse strand $B_-(p)$ were the same and identical to the reference nucleotide $RS(p)$, i.e. for which the condition $B_+(p) = B_-(p) = RS(p)$ held. The result of 71.5% recall is comparable with the fraction of called bases in the reference strain for the *Arabidopsis* project where 79.3% of the queried positions had identical base calls on both strands and the reference sequence. Tables A.1 and A.2 show the base recall for all 20 strains. Already with these simple statistics, wafer-dependent effects are revealed. For instance, we observe that Dom Sufid has more raw base calls matching the reference for three wafers than Tainung 67, although one would expect the opposite judging from the phylogenetic tree and the number of expected SNPs. Moreover, the array on the developmental wafer shows the best quality as the experimental settings were optimised. Due to cost limitations, however, the experimental protocol was modified which decreased the final call rate of the arrays.

In a second step, we investigated the hybridisation performance on a per LR-PCR amplicon and on a per wafer basis. For this assessment, two descriptive measures were used. We first counted the percent of bases called as reference for each LR-PCR amplicon. A base was defined as called analogously to the definition above. Secondly, the mean quality scores for all bases queried by a LR-PCR amplicon were calculated. An average was taken over both the forward and reverse reads. Additionally, we determined the histograms of bases called as reference and mean quality scores across all 20 strains.

We generated figures plotting the fraction of called bases per amplicon in the reference Nipponbare against the fraction of called bases in one of the other 19 strains taking all wafers together and on per wafer basis for wafer 1 to 5. The results are shown for the strains Tainung 67 and SHZ2, a closely related *japonica* strain and a distant *indica* strain, respectively, in the Figures A.1, A.2, A.3 and A.4 in the Appendix. Ideally, one would expect for a closely related strain that most of the data points accumulate on the diagonal line in the upper part of the plot. This is approximately the case for wafer 2, 4 and 5 for strain Tainung 67. Worse performance is observed for wafer 1 and 3 for which the plots show a cluster of data points below the diagonal in the most right part, reflecting that amplicons with a high base call fraction in Nipponbare match lower base call fractions in Tainung 67. The varying performance between the wafers can be explained by both wafer artifacts as well as amplicon failures and unspecific amplification. Another small cluster is observed in the left bottom corner showing amplicons with poor performance in both reference and Tainung 67. This cluster reveals the failures or worse outcomes of the LR-PCRs.

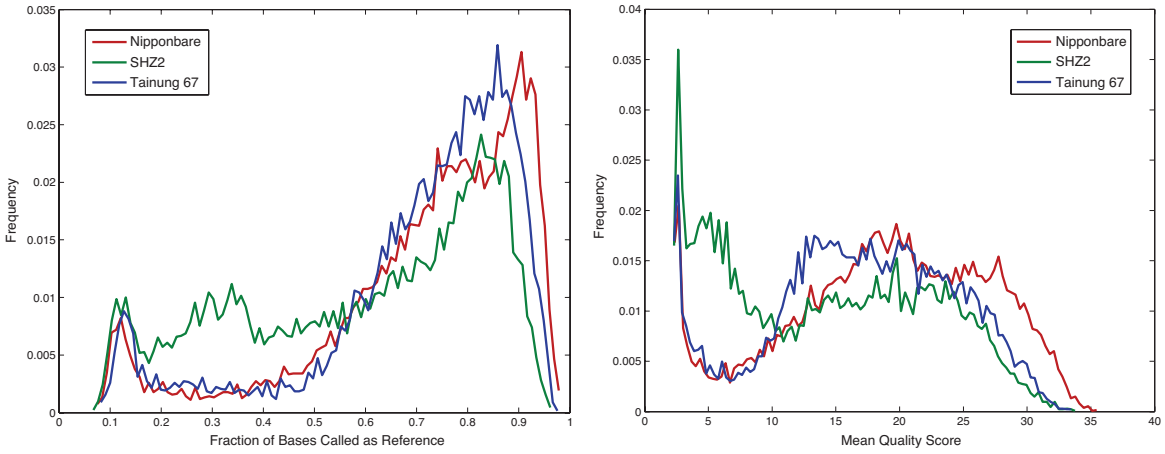
For strain SHZ2 the results of the hybridisation experiments are illustrated in Figure A.3 and A.4. Here, a general decrease in quality is observed. Across all wafers the data points are accumulated below the diagonal line with an obvious shift in wafer 1 and 5 in particular. The worse performance for SHZ2 is also obvious in Table A.1 showing the lowest recall for SHZ2 among all strains. We can explain this striking difference by the its distant relationship to Nipponbare increasing the likelihood for PCR failure and the number of variations in the sequence. This impression is supported by the plots of the other strains (data not shown), observing an inverse correlation between the hybridisation quality and the evolutionary distance

to Nipponbare.

Additionally, we made similar figures showing the mean quality scores on a per amplicon basis. Their interpretation leads to conclusions conform to those we arrived at for the previous investigations, implying a performance dependency on strain and wafer.

Furthermore, we examined the distribution of the observed fractions of bases called as reference and the distribution of the observed mean quality scores. Figure 2.5 shows such histograms for the strain Nipponbare, Tainung 67 and SHZ2. For both *japonica* strains a high number of bases are called. Evidences for amplicon failure can be seen in a second mode of the distribution close to 0% called bases in Figure 2.5(a). The SHZ2 histogram of mean quality scores more clearly illustrates the effect of failures in the PCRs (see Figure 2.5(b)). The histograms for SHZ2 show a more uniform distribution in both measures, lacking a strong separation between low and high recall/quality score values.

Summarising, we conclude from our investigations on the rice hybridisation data that the overall quality is slightly worse than the hybridisation performance in the *Arabidopsis* project. We observed that the hybridisation quality depends on both wafer and strain. Moreover, failures of the LR-PCRs used for the amplification of the target DNA were revealed by histograms applying two measures, showing the drawback of this amplification method. The problem of amplicon failure was not relevant for the *Arabidopsis* resequencing project as the genomic DNA was amplified with random primers, ensuring amplification across the whole genome.



(a) (b)

Figure 2.5: (a) The figure shows the histograms of the fractions of the bases called as reference per amplicon for the strains Nipponbare (red), SHZ2 (green) and Tainung 67 (blue). For both *japonica strains* (Nipponbare and Tainung 67) the number of bases which are called as reference are obviously higher. For SHZ2, lower fractions occur more frequently. (b) The distributions of the mean quality score are illustrated for the three strains on a per amplicon basis. One observes a worse hybridisation performance for SHZ2 compared to the other strains as well. Whole amplicon failures are indicated by the high frequency of amplicons with very low mean quality score.

3 Annotation of Repetitive k-mers in the Reference Genome

This chapter addresses how repetitive k-mers in the reference genome can be efficiently detected. After motivating the need of repetitive a k-mer annotation, we define four types of k-mer matches in Section 3.2 in more detail. We then present the main ideas of our approach and describe implementation details. Finally, the results of the k-mer analysis are shown in Section 3.4.

3.1 Motivation

Resequencing tiling arrays are designed in a way that they can be employed to discriminate single nucleotide polymorphisms. One assumes that each probe hybridises specifically to target DNA from the location corresponding to the probe's origin in the reference genome. However, this assumption is of theoretical nature as the location of the target DNA does not play a role in forming a hybridisation pair, but the thermodynamics of nucleic acid hybridisation is the crucial factor. Thus, a probe binds target polynucleotides such that the formed duplex is energetically optimal. An oligomer occurring several times in the target DNA cross-hybridises with identical probes of other locations, distorting the hybridisation signal. Degenerate repeats are a prominent example causing cross-hybridisation. Figure 3.1 visualises this phenomenom.

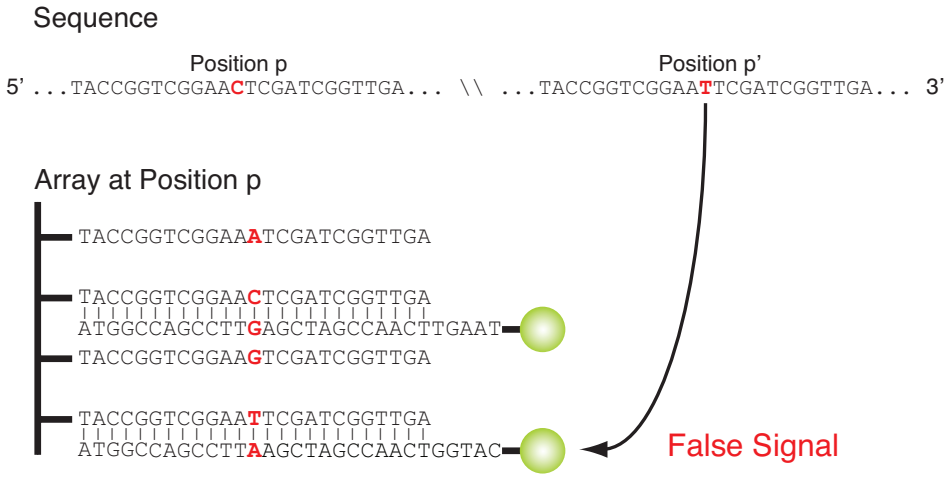


Figure 3.1: The picture illustrates cross-hybridisation. A repetitive 25-mer at position p' with a mutation at the central nucleotide can bind to a probe corresponding to position p and lead to false SNP signal.

The special design of SNP detection arrays, having an octet of probes for each site in the target DNA, increases the chance of cross-hybridisation and therefore unspecific signals. A duplicated oligomer which has a base substitution at its centre position can bind to a non-reference probe of an octet designed for another site (and thus can be mixed up with a SNP signal). Distinguishing this signal from a true SNP signal is very difficult.

This problem appears in a considerable magnitude for natural genomic sequences which are characterised by high repetitiveness. For rice, approximately three fourth of the genome is repetitive. As mentioned in Section 2.2, ∼100 Mb of unique and low copy genomic sequence was used for resequencing. This implies that the resequenced DNA still contains repetitive sequences. It is therefore important to be able to detect oligomers occurring more than once such that repetitive sites can be treated separately in further analyses of the resequencing data.

The goal of the k-mer analysis is the identification of oligomers of length k which match at least to one other k-mer in the target DNA allowing for some degree of degeneracy. These mismatch criteria were set up reasoned by studies on hybridisation energies between non-perfectly matching oligonucleotides. Clearly, it does not matter which base is situated at the centre position of the k-mer as one of the four probes per strand will perfectly bind such a oligonucleotide. Hence, we allow mismatches at the centre base in addition to further mismatches affecting duplex formation.

The thermodynamics of DNA hybridisation is characterised by two main processes. Denaturation, i.e. the disassociation into probe and target molecules, and renaturation, i.e. the forming of the probe-target-duplex, compete with each other in hybridisation (see Burden et al., 2006; Lodish et al., 2004, p. 105/6 Chapter 4). To measure the degree of stability of the duplex, the melting temperature T_m is commonly used as a describing property. T_m is defined as the temperature at which half of the DNA strands are in a double-stranded state and the other half is denaturated. The melting temperature of a certain duplex can be approximated by various physical models. For the comparison of perfect and non-perfect hybridisation duplexes due to a single mutation in one of the strands, Lee et al. (2004) used a model based on nearest-neighbour thermodynamics to calculate the difference ΔT_m between T_m of both duplexes. They investigated the dependency of ΔT_m on the mutation position and showed that mismatches at the end of the duplex have less influence on the duplex stability than at positions close the centre. For this reason, inner one-base mismatches and mismatches at the ends of the oligo-duplexes were considered as separate mismatch types.

In another study by Karaman et al. (2005), hybridisation specificity was examined by comparing relative hybridisation signals from probes perfectly matching the target to those probes with mismatches. They arrived at the conclusion that single base insertions and deletions in probes decrease the hybridisation specificity less than single base substitutions, i.e. one-base bulges in one of the duplex forming oligonucleotides give also rise to oligo-duplexes which are almost as stable as perfect duplexes. Thus, indels of 1 bp length destabilise the oligo-duplex less than single base mismatches do.

Motivated by the results of these investigations, we decided to extend the set of match types used by Georg Zeller in the *Arabidopsis* resequencing project (see Clark et al., 2007) by including so-called bulged matches. Thus, we performed the k-mer analysis distinguishing four match types which are described in more detail in the following section. Note that the definitions of the four match types are exclusive, i.e. the described sets are disjoint.

3.2 Definitions

3.2.1 Exact k-mer Matches

We define all matches at position p as exact k-mer matches which are formed when k-mers from the genomic sequence perfectly hybridise in a duplex with one of the eight array probes interrogating the base at position p. This definition includes k-mers which are identically repeated at some other site in the genome and can thus perfectly hybridised to the non-polymorphic probe in the octet. Moreover, the set of exact k-mer matches contains also those matching oligonucleotide pairs for which a mismatch at position $\lceil \frac{k}{2} \rceil$ is allowed and the bases at all other positions form perfect hydrogen bonds. This corresponds to perfect duplex formation between such a k-mer and one of the three polymorphic probes on the array. It is obvious that also match pairs between a k-mer located on the forward strand and a k-mer belonging to the reverse strand are considered as exact matches. Summarising, an exact k-mer match between a k-mer at position p and a k-mer at position p' can be formally written as:

$$S[p+i] = S[p'+i] \;\; \forall\, i = -\frac{k-1}{2}, \ldots, -1, 1, \ldots, \frac{k-1}{2}$$

where S is the sequence on the forward strand and analogously

$$S[p+i] = \overline{S}[p'+i] \;\; \forall\, i = -\frac{k-1}{2}, \ldots, -1, 1, \ldots, \frac{k-1}{2}$$

where $\overline{S}$ denotes the reverse complemented sequence.

3.2.2 Inexact k-mer Matches

Allowing one mismatch in addition, while ignoring bases at the centre position, we arrive at the definition of inexact k-mer matches. Moreover, the second mismatch position is restricted to the inner parts of the oligonucleotide, excluding two bases at both ends, in order not to overlap with the third match type of short k-mer matches. Thus, we are able to deal with disjoint sets of matches. Formally, the criteria for inexact matches are:

$$S[p+i] = S[p'+i] \;\; \forall\, i \in \left\{ -\frac{k-1}{2}, \ldots, -1, 1, \ldots, \frac{k-1}{2} \right\} \setminus m$$

and

$$S[p+m] \neq S[p'+m]$$

where $m \in \left\{ -\frac{k-1}{2} + 2, \ldots, -1, 1, \ldots, \frac{k-1}{2} - 2 \right\}$.

The total number of allowed possibilities for mismatches in inexact matches is thus $k-5$. In analogy, the criteria hold for match pairs of different strands.

3.2.3 Short k-mer Matches

The definition of short k-mers matches includes all k-mer pairs which have at least one mismatch at the ends. In other words, all bases except the bases at the centre position and at the two positions at the ends have to be identical, requiring at least one end mismatch:

$$S[p+i] = S[p'+i] \quad \forall\, i = -\frac{k-1}{2}+2, \ldots, -1, 1, \ldots, \frac{k-1}{2}-2$$

and

$$\sum_m \delta(S[p+m], S[p'+m]) \geq 1$$

where $m \in \left\{\pm\frac{k-1}{2}, \pm\left(\frac{k-1}{2}-1\right)\right\}$ and $\delta(s_i, s_j) = \begin{cases} 1 & s_i \neq s_j \\ 0 & s_i = s_j \end{cases}$.

Similarly, we define the matching between a forward and a reverse strand.

3.2.4 Bulged k-mer Matches

Bulges in hybridising oligomers are formed when one or more nucleotides remain unpaired. The definition of bulged k-mer matches is restricted to a one-base bulge which is located only on one of the nucleic acid strands. Bulged k-mer matches are then matches between an oligomer of length k and an oligomer of length $k+1$ in which the longer one of the pairing strands contains a bulge of exactly one nucleotide, tolerating mismatches at the centre position.[1] The first and last positions are not included in this analysis as they are not real bulges, but dangling ends which are covered by the definition of short k-mer matches.

Formally, we can describe bulged k-mer matches as:

$$S[p+i] = S[p'+i] \quad \forall\, i < b \text{ and } S[p+i] = S[p'+i+1] \quad \forall\, i \geq b$$

where $i \in \left\{-\frac{k-1}{2}, \ldots, -1, 1, \ldots, \frac{k-1}{2}\right\}$ and $b \in \left\{-\frac{k-1}{2}+1, \ldots, \frac{k-1}{2}\right\}$ under the constraint that the bulge is located in the k-mer corresponding to site p'. Again, the same criteria hold for match pairs of different strands.

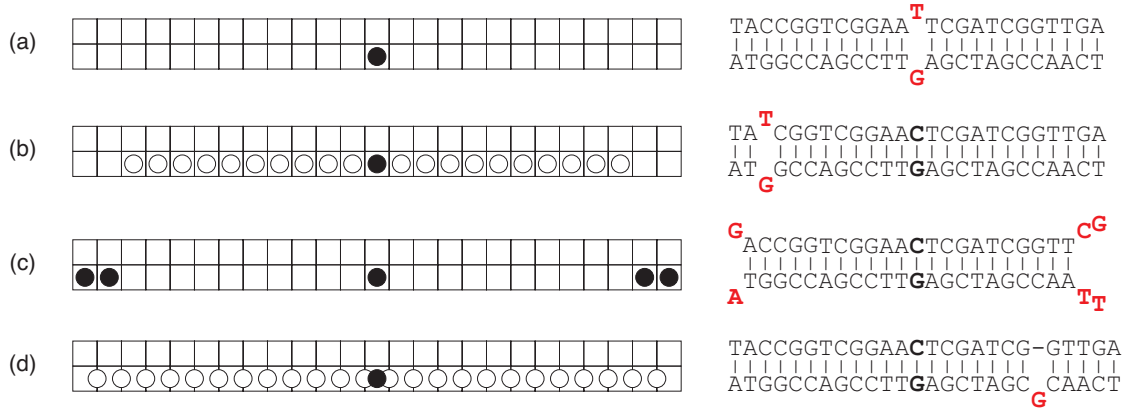


Figure 3.2: The picture shows the tolerated mismatch positions for each of the four k-mer types together with examples on the right-hand side ($k = 25$). (a) Exact 25-mer match. (b) Inexact 25-mer match. (c) Short 25-mer match. (d) Bulged 25-mer match.

[1] In this context, a bulge is defined by position i if it is located between the $i-1$th and ith position of the k-mer.

3.3 Approach

In large parts, the k-mer analyser developed for the 25-mer annotation of the rice genome was based on concepts and implementations by Georg Zeller for the *Arabidopis* project (see Clark et al., 2007, p. S4/5, Supporting Online Material). For rice, the k-mer analyser was ported from Java to C++, adapted to the tiling design for the present resequencing experiments and extended to a version incorporating also bulged k-mer matches.

Briefly, the idea is to generate a list of all 25-mers contained in both the probe DNA, i.e. the DNA which was used for tiling and immobilisation on the arrays, and the amplified target DNA. To each 25-mer its genomic position and affiliation either to a wafer or amplicon were assigned. The list was then sorted to obtain a lexicographically ordered 25-mer list allowing mismatches to some extend. Finally, the sorted list was linearly traversed to report all 25-mer occurring more than once.

In the following, we describe in detail how we realised this idea to compute all exact, inexact, short and bulged 25-mer matches.

3.3.1 Determining the Wafer-amplicon Layout

As indicated in Chapter 2, the tiled sequence lengths summed up to roughly 100 million bp, representing the largely non-repetitive parts of the rice genome. The total length of amplified target sequence exceeded this number since close amplicons were merged due to efficiency reasons and extended to find the optimal PCR primers. Thus, though the tiled sequence and the amplified sequence corresponded in most parts, there was a fraction of positions unique either to the tiled or amplified sequence. Consequently, positions which were only represented by probes on the array, but for which no DNA was amplified in the target pool, could not be queried. This was the case for 20 positions contained in pool 2 and 11, for which all measured signals have to be regarded as noise. Moreover, there were 1,800,000 positions on average per pool for which DNA was amplified, but which were not present as probes on the corresponding arrays. The larger part of these positions were located in one of the 76 amplicons of the developmental pool which were also added to all the other 15 pools.

To assign to each position its affiliation, i.e. whether it was present only in the tiled sequence, unique to the amplified sequence or appeared in both sequences, we determined the blocks of overlapping and unique regions. This information was later used to report only 25-mer matches between tiled and target DNA 25-mers.

3.3.2 Generating a List of All Possible k-mers

As mentioned in Section 2.2, the amplified sequences were split up into pools such that each wafer was divided into three sections, each corresponding to one amplicon pool. Due to this design it sufficed to apply the k-mer analysis pool-wise, i.e. the goal was to find 25-mer matches between the amplified sequence in a particular pool and the complementary probes distributed over the 16 to 17 arrays corresponding to this pool (except for pool 0 for which the amplicons were hybridised to only one array). Per pool roughly 8.5 million bases were investigated (pool 0: 660,000 bases), summing up to more than 127 million bases.

We did not include mitochondrial and chloroplast DNA as done in the k-mer analysis for the *Arabidopsis* project since no whole-genome amplification with random primers was performed, but specific primers were designed to amplify specific regions of the genome. Thus, we assumed that the labelled amplicon pools did not contain sequences amplified from organellar genomes.

For each tiled and/or amplified block each possible $(k+2)$-mer (27-mer in our application with $k = 25$) was determined by shifting a $k+2$ window along the given sequence. We excised one additional base on both ends respectively of the actual k-mer. Due to implementation specific reasons, a two base longer oligomer was needed for the investigation of bulged matches between both strands. Moreover, we accepted only k-mers for which all bases were reliably determined, i.e. which only contained `A`s, `C`s, `G`s and `T`s and no `N`s.

Besides the sequence information, the genome position corresponding to the centre base of the $(k+2)$-mer and the affiliation, i.e. located on a wafer, amplicon or on both, was stored as one so-called tiling object. We also included the reverse-complemented version of the $(k+2)$-mer, denoted then by a negative position value, in order to find k-mer matches between the forward and reverse strand.

On average, 30 million $(k+2)$-mers were generated per pool and 2 million $(k+2)$-mers in the developmental pool 0. Using the encoding which is described in the next paragraph, each tiling object occupied 14 Bytes in memory. This lead to a feasible amount of required memory of 420 Mega Bytes (MB) on average to store a $(k+2)$-mer list per pool.

3.3.3 Encoding k-mer Sequences

If we had stored the DNA sequence straight as an array of type char, we would have needed 27 Bytes to store each $k+2$-mer. Consequently, storing all 30 million 27-mers per pool would have cost 810 MB of memory. As a DNA sequence is limited to an alphabet of only four characters $\Sigma = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ (after discarding ambiguous bases), it can be stored in a more compressed way using 2 Bits per base. Thus, a sequence of length 27 requires only 54 Bits and can be encoded as a 64-Bit long, reducing the memory needed for $(k+2)$-mer storage by 70%. A mapping ϕ from Σ to the set $\Upsilon = \{00, 01, 10, 11\}$ of binary numbers of length 2, encoding each letter of Σ in 2 Bits, was defined as:

$$\phi : \Sigma \longrightarrow \Upsilon$$

$$\text{with } \mathtt{A} \longmapsto 00, \mathtt{C} \longmapsto 01, \mathtt{G} \longmapsto 10, \mathtt{T} \longmapsto 11.$$

The mapping preserves the lexicographical order of the four base characters as a numerical order. By such an encoding, the lower 54 Bits are occupied by the actual encoded 27-mer, wasting 10 Bits of memory per oligomer.[2]

Representing the oligomer sequence as a 64-Bit number simplifies the comparison of two k-mers and makes it more efficient as now only two numbers instead of every single base have to be compared.

[2] Encoding the $(k+2)$-mer in a 64 Bit number, the oligomer length k is bounded to 30.

3.3.4 Comparison-based Sorting of the k-mer List

Efficient sorting is an import step in the k-mer annotation procedure. We therefore used a sorting algorithm which is based on comparison sorts. Introsort combines the benefits of two sorting algorithms, Quicksort and Heapsort. The idea of Quicksort is to sort a list recursively by a three-step divide and conquer strategy (see Cormen et al., 2001, p.145/6, chapter 7). First, an element q is picked to rearrange the list into two partitions, such that one partition contains elements which are less than or equal to q and that each element of the second partition is greater than q. These partitions are then sorted by recursively calling Quicksort. In contrast, Heapsort uses a heap data structure from which the largest value is extracted iteratively from the root position (see Cormen et al., 2001, p. 127-37, Chapter 6). Switching from Quicksort to Heapsort when the recursion depth exceeds a predefined value, Introsort achieves a worst-case runtime of $\Theta(n \log n)$, where n denotes the length of the list to be sorted, and is thus practical for the application to our data.

The advantage of a comparison-based algorithm is the ability to adapt the comparison function in a suitable way. We defined a comparison operation, here a $\leq$-operator, which allowed to mask certain positions of the two compared $(k+2)$-mers, i.e. certain bit pairs in the corresponding encoded $(k+2)$-mers, when sorting. This was realised by the usage of bit masks. For all types of matches, the base at the centre position of the $(k+2)$-mer is not relevant when comparing two $(k+2)$-mers. Thus, the corresponding two bits of the encoded k-mer can be substituted by 00. As we had included one additional base at the both ends, respectively, but only oligomers of length k should be compared, we replaced the two bits encoding the first and the last base of the $(k+2)$-mer by 00. This can be easily done by AND-masking using a mask with two zeros at the sites to be tolerated and ones elsewhere. This concept was extended to also incorporate tolerated bases at other positions for inexact and short matches by adding 00 in the mask at sites corresponding to mismatch positions.

Incorporating bulged k-mers was more challenging: A given position i at which a base is located in a bulge was erased in the $(k+2)$-mer using AND-masks. If $(k+2)$-mer came from the forward strand, the bits were shifted by two positions in low-bit direction; otherwise a shift of two positions in high-bit direction was applied as part of the comparison operation. The masking and erasing operations are illustrated in Figure 3.3.

As we used individual masks and the erase function to generate k-mers during comparison, the content of $(k+2)$-mer list remained static. Thus, we did not need to recompute the list after sorting with a particular mask and could reuse the list for computing all types of k-mer matches.

position		1	2	3	4	5	6	7		
1st 7-mer κ	T	C	T	C	G	A	T	C	G		
encoded	11	01	11	01	10	00	11	01	10	≡	226,870
(a) Exact match pair											
κ masked	00	01	11	01	00	00	11	01	00	≡	29,748
2nd 7-mer	G	C	T	C	A	A	T	C	T		
encoded	10	01	11	01	00	00	11	01	11	≡	160,823
masked	00	01	11	01	00	00	11	01	00	≡	29,748
(b) Inexact match pair											
κ masked	00	01	11	00	00	00	11	01	00	≡	28,724
2nd 7-mer	C	C	T	G	T	A	T	C	C		
encoded	01	01	11	10	11	00	11	01	01	≡	97,077
masked	00	01	11	00	00	00	11	01	00	≡	28,724
(c) Short match pair											
κ masked	00	00	00	01	00	00	11	00	00	≡	1,072
2nd 7-mer	G	G	G	C	G	A	T	G	A		
encoded	10	10	10	01	10	00	11	10	00	≡	173,624
masked	00	00	00	01	00	00	11	00	00	≡	1,072
(d) Bulged match pair											
κ masked	00	01	11	01	00	00	11	01	00	≡	29,748
2nd 7-mer	T	C	T	C	G	A	A	T	C		
encoded	11	01	11	01	10	00	00	11	01	≡	226,829
erased	T	C	T	C	G	A	T	C			
masked	00	01	11	01	00	00	11	01	00	≡	29,748

Figure 3.3: The comparison between the 7-mer κ and a second 7-mer with the usage of bit masks for the four match types is shown. Each 7-mer is illustrated as DNA sequence and as the corresponding encoded bit string. Positions of interest are coloured in dark grey. To facilitate the comparison between the encoded 7-mers, we added the corresponding numbers in decimal representation in the right column.

3.3.5 Reporting k-mer Matches

To report all pairs of matching position (p, p'), the sorted $(k+2)$-mer list was traversed using two pointers b and e defining the begin and end of a set of matching entries. First, b was set to the first list entry and e was advanced from the ith to the $i+1$th entry beginning at b as long as the ith and the $i+1$th entry were equal. If both differed, b and e defined the maximal interval of equal entries. If b was less than e, the set of all possible allowed match pairs between the entries in the interval $[b, e]$ was reported. The matching was restricted to pairs of the members with different affiliation, i.e. requiring that one k-mer originated from the wafer sequence, whilst the other one is located on the amplicon sequence. Then b and e were advanced to point to the next entry. In that way, the list was processed until e reached the end of the list.

Sorting and reporting matches had to be done once to detect exact and 20 times to inexact 25-mer matches with mismatches at positions $i = 3, \ldots, 11, 12, 14, 15, \ldots, 23$. Additionally, we had one mask to determine short matches. Bulges were permitted to appear at all positions except the first and the last positions and we thus applied the sorting and reporting procedure 24 times. In total, sorting and reporting was performed 46 times, taking at most 85 minutes for each pool.

3.4 Results and Discussion

3.4.1 Postprocessing of Bulged k-mer Matches

After computing the pairs for exact, inexact, short and bulged 25-mer matches, we processed the bulged 25-mer data such that only match pairs (p, p') satisfying $|p - p'| \geq 25$ were included in statistics and further investigations. This restriction was motivated by the observation that poly`N`-regions, i.e. regions consisting throughout of either `A`s, `C`s, `G`s or `T`s, contribute heavily to the number of bulged 25-mer matches. This can be easily understood by the following simplified example: Given a 9-mer κ_p `AAAATCGAC` with a poly`A`-region of length 4 at the beginning and corresponding to some position p in the genome, the k-mer analysis of bulges results in three bulged matches to the 9-mer κ_{p+1} at position $p+1$ (where the nucleotides at the second, third and fourth position of κ_p form a bulge). To get a feeling of the magnitude of the poly`N`-regions fraction, we traversed the rice genome determining the number of poly`N`-regions of length n, $n > 1$. We discovered that 47% of the nucleotides in the genome were followed by at least one identical nucleotide. As we could not discard half of the resequenced positions for SNP prediction, we decided to exclude bulged k-mer matches which arose from poly`N`-regions.

3.4.2 Final Results

Taking the 25-mer matches from all match types and all pools together, the k-mer analysis resulted in 17,493,032 match pairs. Results are presented for each match type individually in Table 3.1. Table A.3 in the Appendix shows the distribution of k-mer matches over the different pools.

More interesting and meaningful than looking at the statistics of matching k-mer pairs is the investigation of positions which were interrogated on the arrays and were not unique. According to our k-mer analysis based on the pool design, a total of 5,160,864 positions are repetitive, making up 5.16% of all positions used for tiling. Table A.4 in the Appendix illustrates the counts for all individual pools and global counts.

Table 3.1: Repetitive 25-mer matches for *O. sativa japonica* cv. Nipponbare. The percentage of repetitive positions is given in squared brackets.

25-mer Match Type	Match Pairs	Repetitive Positions
Exact	3,002,679	2,528,180 [2.53]
Inexact	6,027,990	2,530,991 [2.53]
Short	2,478,436	940,055 [0.94]
Bulged	5,983,927	1,569,404 [1.57]
Union	17,493,032	5,160,864 [5.16]

3.4.3 Analysis of the k-mer Data

With the annotation of the repetitive sites in the tiled sequence, we were then able to examine the effect of non-unique 25-mers on the hybridisation data. For that purpose, we determined the densities of both maximal hybridisation intensities[3] and maximal quality scores over both strands, first taking data from all positions with unique probes and secondly from all repetitive positions. Comparing these histograms, we could characterise the difference between the two distributions.

We first looked at the effects of repetitive sites which had exact 25-mer matches to other positions (see Figure 3.4). In Figure 3.4(a) we can identify an obvious shift of the intensity distribution towards higher intensities for repetitive positions. This shift confirms the assumption that repetitive sites effect the actual measured hybridisation intensities by contributing to the set of possible target DNA and thus increasing the hybridisation signal. We further investigated the effect on the distribution of the maximal quality scores. Figure 3.4(b) shows no significant dependency of the distribution on whether the sites are unique or repetitive. This results meets our expectations as exact 25-mer matches bind perfectly to one of the probes of the octet. Thus, the quality score is not affected, masking sites of potential spurious SNP calls. The small increase in frequency of the low quality scores at non-repetitive sites is likely to occur due to random effects.

Furthermore, we performed the same analyses for short 25-mer matches. The results are given in Figure 3.5. The two intensity histograms are about as much similar as those plotted for the exact 25-mer matches. As the intensity histogram of repetitive positions is shifted towards higher intensities as observed for exact 25-mer matches, we can assume that the stability of the oligo-duplex between short 25-mer matches is comparable to the one for perfect

[3] Note that intensities were plotted on a logarithmic scale.

matches. We separately analysed the distributions depending on the mismatch positions. Here, repetitive positions with at least one mismatch at a outer, i.e 1st or 25th, site of the 25-mer were distinguished from those with mismatches at the second and the second last site (figures not shown). We observed no significant differences between these two mismatch positions reflecting that the location of the mismatch does not appreciably affect the intensity distributions with respect to short 25-mer matches. We also studied the histograms of the maximal quality scores for this match type. Now a difference between the two histograms is observed. The density of low quality scores (between 1 and 5) of probes with matches to targets of other locations is increased, whereas there are fewer probes with extremely high quality scores compared to non-repetitive positions.

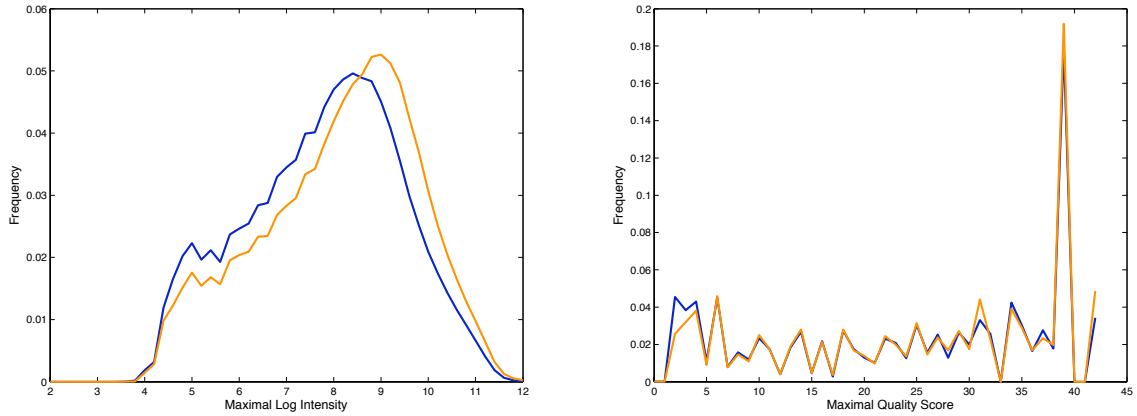
Concerning inexact 25-mer matches, we also performed a mismatch position-dependent analysis of the distributions. The results are given for maximal intensity histograms in Figure A.5 and for maximal quality score histograms in Figure A.6. For mismatch positions located at the ends of the 25-mer, i.e. mismatches at position 3, 4, 5 and 21, 22, 23, we observe intensity distributions similar to those for short 25-mer matches. The closer the mismatch position is to the centre position, the more the distribution of the repetitive positions is shifted towards lower intensities. A possible explanation is that mismatches close to the centre position of the oligo-duplex affect its stability more strongly than mismatches at the ends. Hence, the intensities of these duplexes are on average lower than the intensities of perfect duplexes. The histograms of the maximal quality scores show reduced frequency of higher quality scores accounting for an increase in the frequency of lower quality scores, illustrating that mismatched base pairs in a DNA duplex affect the ability of calling the true base.

Finally, distributions were compared for bulged 25-mer matches. Taking all bulge positions together, no significant difference between the distributions was observed. Analysing the distributions by breaking them down into the different bulge positions, we can detect position-dependent differences (see Figures A.7 and A.8). Duplexes with bulges in outer regions result in lower intensities. Moreover, the intensities are distributed more evenly over a log intensity range of 4 to 9, in contrast to non-repetitive sites for which the density curves exhibit a clear peak at an intensity of $\sim$8.5. A possible explanation for this phenomenom could be that duplexes containing bulges at the ends are energetically unfavourable, thus resulting in lower hybridisation intensities.

The opposite effect can be observed for intensity distributions at sites for which 25-mer matches with bulges near the centre positions of the 25-mer were found. Here, we notice a similar trend compared to exact 25-mer matches with an even more extreme shift towards higher intensities. This implies that duplexes containing bulges close to the centre behave similarly to perfect matches. The extreme dependency on the bulge location is also observed in the quality score histograms. The more central the bulge is located, the more likely an error in base calling occurs as a consequence of the distortion introduced by bulged 25-mer matches. We find a dramatic quality score reduction for central bulges. This is probably caused by reduced peak separation due to lower sequence-specificity at positions in the duplex close to the bulge, i.e. a 25-mer with a bulge at the centre position binds to all quartet probes similarly.

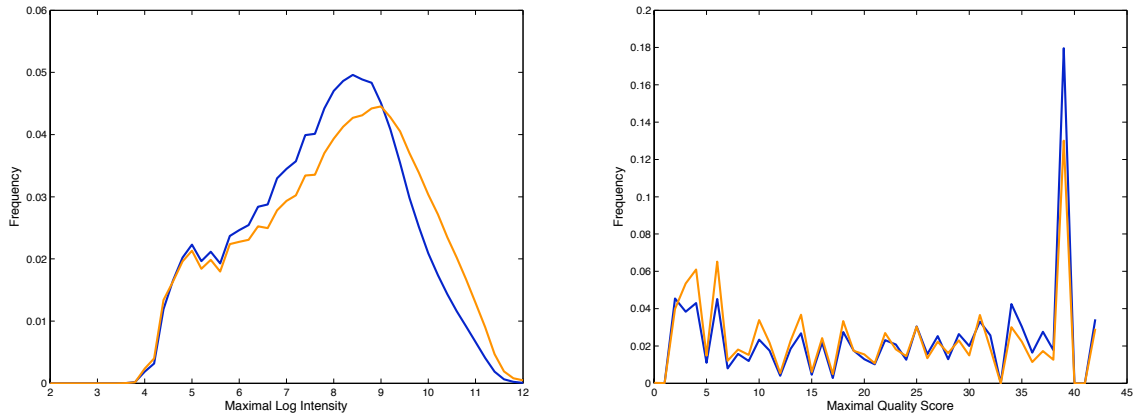
Summarising our observations and conclusions, we showed that repetitive 25-mers do have an effect on the hybridisation intensities in general. Mismatches near the centre position destabilise the duplex more strongly, i.e. hybridisation depends on the location of the mis-

match. A position-dependency of the hybridisation energy can also be stated for bulged 25-mer matches. Furthermore, 25-mer with indels probably form a more stable duplex than non-perfect matches do. Our observations are thus consistent with the results of the experiments done by Karaman et al. (2005).



(a) (b)

Figure 3.4: Effects of exact 25-mer matches on hybridisation. (a) The distribution of the maximal hybridisation intensities are illustrated in blue for all unique positions and in orange for positions with exact 25-mer matches. (b) The histograms of maximal quality scores are shown for positions with (orange) and without (blue) exact 25-mer matches.



(a) (b)

Figure 3.5: Effects of short 25-mer matches on hybridisation. (a) The distribution of the maximal hybridisation intensities are illustrated in blue for all unique positions and in orange for positions with short 25-mer matches. (b) The histograms of maximal quality scores are shown for positions with (orange) and without (blue) short 25-mer matches.

3.4.4 Application of the k-mer Data

As mentioned in the introduction of this chapter, repetitive k-mers are critical as they interfere with SNP identification leading to false predictions. They can cause false signals by either suppressing a true SNP signal or generating an incorrect SNP signal. A k-mer which occurs several times in the genome and is identical to the reference probe at position p suppresses a possible SNP signal at that site. This leads to an increase in the number of false negatives and thus decreases sensitivity. When a repetitive k-mer mimics a SNP, the number of false positives is increased which reduces specificity.

False positives are most likely to be observed and of consequence at positions where the number of the nucleotide at the centre position of the repetitive k-mers exceeds the counts of matches supporting the reference nucleotide. We identified these so-called dominating 25-mer positions as a subset of the positions with repetitive 25-mers. The results are given in Table 3.2. Information on dominating 25-mer positions was also used in the learning algorithm described in Chapter 5. Moreover, we investigated the influence of dominating and repetitive 25-mers on the performance of the two SNP calling methods (see Section 4.3 and 5.2.1).

Table 3.2: Dominating repetitive 25-mer matches for *O. sativa japonica* strain Nipponbare. The percentage of repetitive positions is given in squared brackets.

25-mer Match Type	Match Pairs	Repetitive Positions
Exact	1,664,629	1,537,687 [1.54]
Inexact	212,514	208,776 [0.21]
Short	663,823	610,036 [0.61]
Bulged	1,406,553	1,048,578 [1.05]
Union	3,947,519	2,693,837 [2.69]

4 The Dideoxy Sequencing Data

The goal of this thesis was to come up with a method for SNP prediction based on a supervised learning method. Such methods require labelled examples as a training set. For this purpose, we compiled a set of true SNP and non-SNP positions generated from sequences obtained by dideoxy sequencing.[1] As complex models such as Support Vector Machines with radial basis function kernels profit from a large label set, we ensured that we sequenced a reasonable number of polymorphic positions.

The aim of this chapter is to give a description of the dideoxy sequencing data we utilised for training of the machine learning algorithms and for quality assessment of our polymorphism predictions. We start by illustrating how we selected the samples for sequencing and how we processed the raw data. The bigger part of this chapter focuses on the performance analysis of a model-based SNP calling method and on assessing the quality of the dideoxy sequencing data.

4.1 Sample Selection

The *Arabidopsis* resequencing project benefited from a data set previously sampled by dideoxy sequencing: Nordborg et al. (2005) resequenced 876 fragments of average length of 583 bp spaced throughout the genome in a sample of 96 *Arabidopsis thaliana* individuals. The alignments of the sequenced fragments to the reference strain were manually curated. Each potential SNP was double checked by visual inspection of the sequencing chromatograms. This careful proceeding resulted in a highly reliable data set of confirmed polymorphisms. A subset containing sequenced data from 19 *Arabidopsis* strains was applied in the *Arabidopsis* resequencing project, identifying between 2,266 and 3,949 single nucleotide polymorphisms per strain relative to the reference (see Clark et al., 2007).

In the case of the rice resequencing project, such a curated data set of polymorphisms was not previously available. We thus had to select fragments for sequencing ourselves. Based on the *Arabidopsis* data we simulated two different scenarios. First, the recall was determined as a function of the number of sequenced strains, i.e. how often a particular position was sequenced for different strains, using all positions at once. Secondly, we simulated SNP calling for various numbers of sequenced positions each one sequenced exactly for only strain and calculated the recall rates. The results are given in Figure 4.1.

Due to cost limitations, the number of possible positions to be sequenced was restricted to roughly 2 Mb, enabling to sequence approximately 3,000 fragments in total. Adjusting for the different polymorphism level in rice (with ∼3-fold lower SNP density), we estimated that a 3-fold number of fragments had to be sequenced for rice compared to *Arabidopsis* to obtain a comparable number of SNPs. One possibility was to sequence the set of 2 Mb

[1] Dideoxy sequencing was performed at The Institute for Genomic Research, USA.

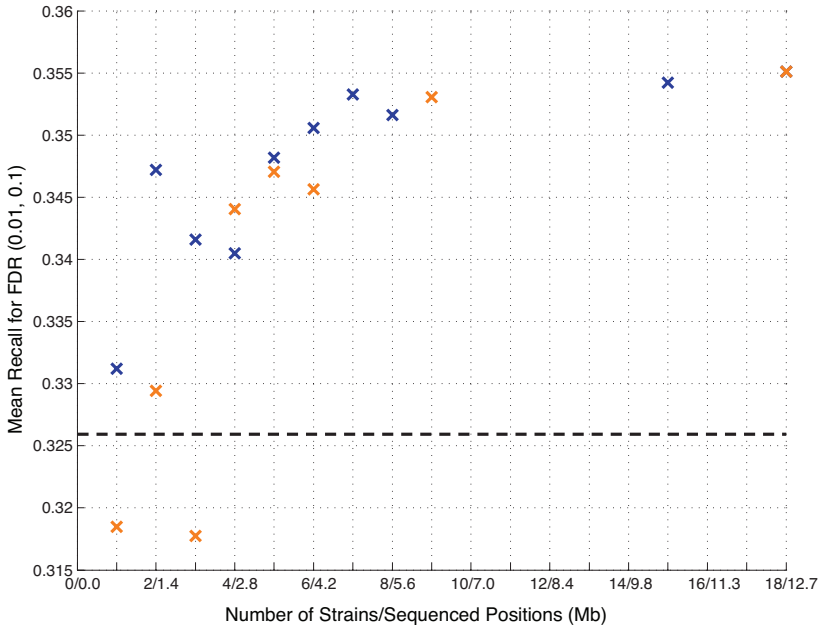


Figure 4.1: Simulations of potential recalls depending on the selection strategy for dideoxy sequencing. The blue crosses denote the recall values when each position was sequenced exactly for a single strain. The other simulation is illustrated in orange crosses. Here, each position was sequenced for a defined number of strains, increasing the number of strains along the x-axis. The dashed line denotes the recall achieved for the machine learning method with training on a per strain basis in the *Arabidopis* project. The figure shows that for fewer sequenced positions the strategy of having positions unique to one strain as a training set results in a better performance. Furthermore, one can observe that recall improves only marginally when more than ten strains are sequenced at every position. The plot was made by Gabriele Schweikert.

positions for every strain, resulting in ∼100,000 unique positions and thus requiring only 100,000 primer pairs. Alternatively, we could sequence ∼2,000,000 positions each for a single strain, only resulting in 3,000 fragments, but requiring obviously more primers. Although the second design seemed to be potentially more expensive, it turned out that both designs would require the same amount of primers in pratice due to experimental design of having primers already provided on 96-well plates. Furthermore according to our simulations, a set which contains more positions unique to one strain would result in a higher recall. Taking all these considerations together, we chose the second strategy and sequenced positions unique to a single strain with the additional constraint that the whole fragment had to be tiled on the array. We sampled the fragments randomly across the genome to avoid a bias towards certain region, e.g. towards coding regions.

Due to the experimental design and cost issues, we chose 96 amplicons to be sequenced for nine strains closely related to the reference Nipponbare (Azucena, Cypress, Dular, LTH, M 202, Moroberekan, Nipponbare, Sadu Cho and Tainung 67) and 192 amplicons for 11 more distantly related strains (Aswina, Dom Sufid, FR 13A, IR 64, Minghui 63, N 22, Pokkali, Rayada, SHZ 2, Swarna and Zhenshan 97). Because of the evolutionary distance to the reference Nipponbare, more SNPs were expected for the more distantly related strains. We

therefore sequenced a higher number of fragments for those strains. In total, 2,976 amplicons were sequenced.

In a second step, additional dideoxy sequencing was performed in order to have sequence information for a subset of fragments not only for a single strain, but also for two additional strains which are closely related to the strain for which the fragment had already been sequenced. As we were constrained by the number of possible experiments, we excluded the reference strain Nipponbare and three other strains (LTH, M 202 and Tainung 67) for which the polymorphism level is expected to be low. Moreover, at most 96 fragments for each previously sequenced strain were resequenced for the two other strains. For the second round, 1,536 fragments were used for sequencing. Having then sequenced each selected fragment three times for different, but closely related strains, we could reliably determine SNPs from the dideoxy sequence traces. In ∼70% of all chosen fragments, PCR was successful and the sequencing resulted in a reasonable sequence, yielding 3,161 sequenced amplicons in total.

4.2 Processing the Raw Data

Sequencing was conducted by Kevin Childs and Robin Buell at The Institute for Genomic Research (TIGR), USA. Kevin Childs was also responsible for aligning the sequence data from forward and reverse read to the corresponding region in the reference genome.

The fragments of interest were amplified by PCR with primers designed from the Nipponbare reference sequence and sequenced in both directions using the Sanger dideoxy sequencing technique. Based on the sequence chromatograms base calling was performed together with the assignment of a quality score reflecting the error probability of the base call. These initial sequences were then trimmed to remove sequencing primers and regions of low quality. For those products which were successfully sequenced for both the forward and the reverse read, the two reads were assembled using Phrap,[2] a programme for shotgun DNA sequence assembly. In parallel, the quality score for each base of the consensus sequence was computed by combining the scores of the two reads if they overlapped, otherwise the quality score of the corresponding single read was taken. ClustalW (Thompson et al., 1994) was applied to align the consensus sequence to the reference sequence.

From these alignments, we parsed the information on single nucleotide substitutions, insertions and deletions. We excluded fragments for which the alignment to the reference sequence resulted in a very high (more than 50) and hence not reliable number of SNPs. After the preprocessing step, the dideoxy sequencing set comprised 1,755,395 positions across all strains of which 9,499 positions were assigned to be polymorphic. Table 4.1 gives an overview over the number of SNPs detected by dideoxy sequencing and shows their distribution over the different genomic regions.

[2] http://www.phrap.org

Table 4.1: The number of SNPs for the different sequence types in the dideoxy sequencing set across all strains grouped by the five subpopulations. The fraction to all SNPs per strain are given in squared brackets.

Strain	All SNPs	Coding	UTR	Intron	Intergenic
LTH	42	3 [07.1]	7 [16.7]	5 [11.9]	27 [64.3]
M 202	55	12 [21.8]	1 [01.8]	1 [01.8]	41 [74.6]
Tainung 67	9	2 [22.2]	4 [44.4]	2 [22.2]	1 [11.1]
Azucena	222	28 [12.6]	20 [09.0]	20 [09.0]	154 [69.4]
Cypress	412	24 [05.8]	40 [09.7]	46 [11.2]	302 [73.3]
Moroberekan	253	6 [02.4]	12 [04.7]	40 [15.8]	195 [77.1]
Dom Sufid	82	12 [14.6]	10 [12.2]	20 [24.4]	40 [48.8]
Dular	654	100 [15.3]	74 [11.3]	73 [11.2]	407 [62.2]
FR 13A	688	97 [14.1]	122 [17.7]	84 [12.2]	385 [56.0]
N 22	854	127 [14.9]	93 [10.9]	79 [09.3]	555 [65.0]
Rayada	903	94 [10.4]	123 [13.6]	90 [10.0]	596 [66.0]
Aswina	414	54 [13.0]	43 [10.4]	67 [16.1]	250 [60.4]
IR 64	599	51 [08.5]	116 [19.4]	126 [21.0]	306 [51.1]
Minghui 63	797	96 [12.1]	99 [12.4]	75 [09.4]	527 [66.1]
Pokkali	908	139 [15.3]	131 [14.4]	104 [11.5]	534 [58.8]
Sadu Cho	414	113 [27.3]	38 [09.2]	46 [11.1]	217 [52.4]
SHZ2	685	125 [18.3]	83 [12.1]	80 [11.9]	397 [58.0]
Swarna	924	128 [13.9]	106 [11.5]	91 [09.9]	599 [64.8]
Zhenshan 97	584	67 [11.5]	106 [18.2]	75 [12.8]	336 [57.5]
Total	9,499	1,278 [13.5]	1,228 [12.9]	1,124 [11.8]	5,869 [61.8]

4.3 Quality Assessment of MB SNP calls

With the first subset of amplicons sequenced, we performed simple analyses to check the performance of the SNP classifier. At that time, we had only access to SNP predictions made by Perlegen Sciences using a model-based (MB) approach. Thus, all performance analyses in this section refer to Perlegen's SNP calling method.

The MB method was developed for SNP discovery using array-based resequencing and has been applied in the human (Hinds et al., 2005), *Arabidopsis* (Clark et al., 2007) and mouse (Frazer et al., 2007) resequencing projects. We just briefly describe the method to give an idea of the main concept, more details are found in Clark et al. (2007, p. S5/6, Supporting Online Material). The method is based on both intensity measurements and raw base calls. It generates alternate consensus calls for raw base calls which differ from the reference and have significantly higher intensity (averaged over both strands) than vicinal calls in a certain window around the site of interest. Putative polymorphic sites have to pass an additional test including adjustments for differences in brightness between array scans. When a SNP

is found in at least one strain, the base calling criteria is relaxed for other strains allowing alternate calls confirmed on only one strand.

To investigate the accuracy and quality of the SNP detection methods, we used two measures. The false discovery rate (FDR) measures the fraction of spuriously predicted positives relative to all predicted positives (PP), i.e. how often the predictor is wrong when it calls a SNP:

$$\mathrm{FDR} = \frac{\mathrm{FP}}{\mathrm{PP}} = \frac{\mathrm{FP}}{\mathrm{TP} + \mathrm{FP}}$$

where TP denotes the number of true positive predictions and FP the number of false positive predictions. The fraction of true positives and positives (P), i.e. true SNP positions which are recovered by the predictor, is denoted by recall or sensitivity:

$$\mathrm{recall} = \frac{\mathrm{TP}}{\mathrm{P}} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

We first determined recall and FDR for all strains together and secondly for each strain separately. Furthermore, we performed the same analyses on a per wafer basis and distinguished between the sequence type. The performance measurements were also calculated excluding positions with repetitive 25-mer matches which were determined by the method described in Chapter 3. The results are given in Table 4.2 and Table A.5 in the Appendix.

Unexpectedly, in terms of both recall and FDR the performance of the MB method was considerably worse than the performance obtained for the MB method applied to the *Arabidopsis* resequencing data (see Clark et al., 2007, Table S3, Supporting Online Material). Excluding repetitive 25-mer positions from performance analyses, we observed that the FDR improved from 16.0% to 14.9%. With the filtering, we did not only reduce the number of false positives, but also true positives were filtered out. Consequently, the recall decreased when filtering for 25-mer repetitive probes from 12.2% to 11.8%.

For the *Arabidopsis* strains, the recall ranged between 16.2% and 23.1%, whereas we observed variations in the recall values of the different rice strains from 5.9% to 22.2% with an average of 12.2%. The performance difference between the two species was shown in their highly varying FDRs in particular: the FDR of the rice MB predictor was 16.0% on average, ranging from 0% to 61.5% in contrast to FDR values from 0.7% to 4.1% for the *Arabidopsis* strains. Best recall rates were achieved for the *aus* strains. The observed variations between the different subspecies are likely to be due to differing numbers of labelled examples.

In the analysis on a per wafer basis, we observed a better performance for wafer 2 and 3 than for the other three wafers, indicating systematic problems due to experimental artifacts.

We also looked at the performance evaluated separately for coding, UTR, intronic and intergenetic sequences. The highest recall was achieved for SNP positions in coding regions. For other sequence types, recall was significantly reduced. The MB recall for *Arabidopsis* showed a similar trend towards higher values in coding regions. An even more extreme difference between the four sequence regions was observed ($\sim$2.5 fold higher recall of coding to other regions in *Arabidopsis*, compared to a $\sim$1.5 fold in rice).

Table 4.2: The number of SNP calls predicted by the MB method are shown with the two performance measures (recall, FDR) in %.

	MB SNP Calls	Recall	FDR
Total	1,385	12.2	16.0
Wafer			
1	232	12.7	19.8
2	329	13.9	12.8
3	290	13.6	12.8
4	248	09.9	16.5
5	286	11.4	19.6
Sequence Type			
Coding	291	19.1	16.2
UTR	187	13.0	15.0
Intron	187	13.0	21.9
Intergenic	720	10.5	14.7

4.3.1 Performance Dependency on Hybridisation Quality Scores

Finally, we investigated the performance dependency on hybridisation quality scores. For this purpose, we looked at the variation of recall and FDR filtering on the prediction level with respect to the hybridisation quality score. We determined the number of true and false positives for which the hybridisation quality scores were greater than a certain threshold with values ranging from 0 to 90. The curves for the resulting recall and FDR at increasing quality score thresholds are illustrated in Figure 4.2. We also calculated the distribution of the hybridisation quality scores at the MB SNP calls showing the magnitude of remaining SNP calls after thresholding. As we expected, the recall decreases with increasing threshold since the number of true positives is reduced, but the total number of SNP positions assigned by dideoxy sequencing is invariant. In contrast to that, the FDR decreases to 8% until a quality score threshold less than 25 is reached and keeps roughly the same for higher scores. The striking peak for a threshold of 40 is likely due to the small number of true positive at that level. From this experiment we can conclude that a filter excluding positions with hybridisation quality score less than 8 is feasible to reduce the FDR by one percentage point while loosing a minimal number of true positives.

As the evaluation results based on the MB method in comparison to the performance in the *Arabidopis* project were appalling poor, we did further analysis to locate possible reasons explaining the observed performance.

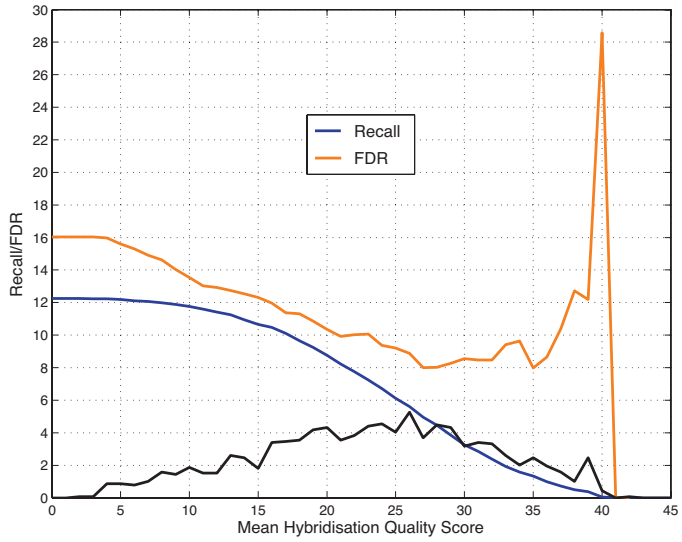


Figure 4.2: MB performance dependency on hybridisation quality scores. The figure shows the distribution of hybridisation quality scores (black), the recall (blue) and the FDR (orange) as a function of the mean hybridisation quality score cut-off.

4.4 Quality Assessment of the Dideoxy Sequencing Data

4.4.1 A Problem with Sample Contamination?

To determine whether the SNP information extracted from the dideoxy sequencing data was reliable, we examined a subset of false positives in more detail. For this purpose, we visualised the maximal hybridisation intensities averaged over both strands of the target strain and Nipponbare, respectively, and the mean intensities of the probes corresponding to the reference base for all fragments in the selected subset. Additionally, we highlighted SNP positions determined by dideoxy sequencing and positions predicted as SNPs by the MB method. Comparing the patterns for the different fragments, we could not reveal a clear evidence for sample contamination.

The only striking observation was the following: In one fragment sequenced for the *indica* strain Aswina, three SNP positions predicted by the MB method could be identified by an obvious SNP pattern in the intensity profiles, but no evidence for any SNP was given by the dideoxy sequence (see Figure 4.3). We therefore suspected that the sequence determined by PCR and dideoxy sequencing was not identical to the sequence which was hybridised to the array. Using BLAT (Kent, 2002) to align the dideoxy fragment to the reference genome of the *indica* subspecies 93-11, we detected that the dideoxy sequence differs in exactly these three positions from 93-11. This is surprising as Aswina itself belongs to the *indica* subspecies and is therefore expected to be more similar to 93-11 than to Nipponbare. We concluded that in this case a false fragment was amplified and sequenced either due to contamination by a *japonica* strain or sample swapping. Motivated by this observation, we then checked all other fragments which had been sequenced at that point for sample swapping by aligning them both to the *japonica* and *indica* reference. If the number of sequence differences to the *indica* reference was significantly higher than to *japonica* reference for a fragment assumed to be sequenced from an *indica* strain or vice versa, we assumed a sample swapping for these cases.

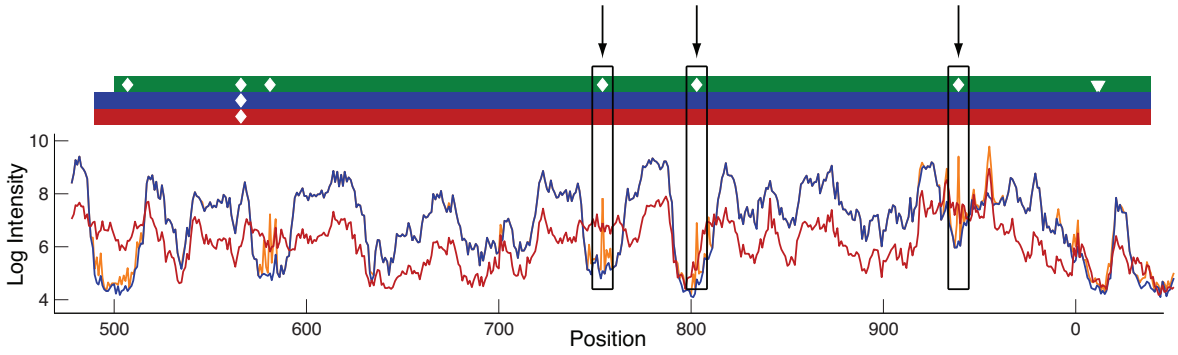


Figure 4.3: The figure shows intensity patterns and alignment results for a fragment ranging from 30,584,490 to 30,585,039 on chromosome 4 sequenced for the strain Aswina. The maximal hybridisation intensities averaged over both strands for Aswina (orange) and Nipponbare (red), respectively, and the mean intensities of the probes corresponding to the reference base for Aswina (blue) are visualised. The bars denote sequence information determined by dideoxy sequencing (red), the alignment against the *japonica* reference (blue) and the alignment against the *indica* reference (green). SNPs are represented by white diamonds and insertions by white triangles. The black boxes denote SNP predictions by the MB method, also showing an obvious SNP signal in the intensity patterns. Note that the x-axis only shows the last three digits of the positions.

Potential contamination with DNA from other strains could definitely be detected by these investigations only in a few cases for the strain Aswina. Altogether, we could not observe a pattern of systematic sample swapping.

To further examine potential contamination, we manually checked dideoxy sequencing trace files of sequences corresponding to a subset of false positives. If the dideoxy samples were contaminated with sequences from the reference or another strain which was not polymorphic at the site of interest, the reference signal could suppress the SNP signal. Furthermore, we asked whether potential SNPs were missed due to heterozygosity of the strain leading to similar high intensity peaks of the two allelic bases. Looking at the trace files, Detlef Weigel observed a clear reference call in the corresponding trace file at almost all of the investigated positions where the MB method called a SNP. Moreover, we could not observe potential heterozygote calls either. From these observations, we concluded that sample contamination and heterozygosity did not cause false base calls in the dideoxy sequences.

In addition, we looked at 11 positions for which a non-reference call was predicted from the dideoxy sequencing for the reference Nipponbare. Looking at the 11 SNPs in Nipponbare assigned by dideoxy sequencing it turned out that these bases were probably false calls. All the 11 cases seemed to be due to problems in the base calling step. In contrast to that, false base calling could not be observed investigating the sequencing trace files of false positive MB SNP calls for other strains.

At that point, we concluded that our dideoxy sequencing data was of acceptable quality. To have further confirmation, we then decided to measure the MB performance using another set of dideoxy sequences.

4.4.2 Usage of a Second Set of Dideoxy Sequences

Fortunately, we had access to a data set used for a genome-wide polymorphism study in rice recently published by Caicedo et al. (2007). The authors of this paper investigated DNA sequence variations between rice and its wild ancestor having sequenced 111 randomly chosen gene fragments across 72 domesticated and 21 wild rice strains. Three of the domesticated rice strains were contained in our set of 20 strains, including Azucena, Dom Sufid and IR 64, each representing one of the main *Oryza sativa* subspecies *japonica*, *aromatic* and *indica*. We processed the selected data in order to map the coordinates based on the TIGR genome release to coordinates based on the IRGSP genome release and to extract the positions for which probes on the tiling arrays and thus hybridisation measurements existed. In total, we obtained 193,281 positions including 387 SNPs (0.2%). This set was used to validate the overall performance of the MB method, also on a per sequence type basis. The results are given in Table 4.3. The FDR is of similar magnitude as the FDR observed on our dideoxy set. The increased overall recall can in part be explained by the fact that the sequences chosen for dideoxy sequencing were sampled from genes, i.e. one expects a higher recall for SNP positions in coding regions according to the observation made in the *Arabidopsis* project. Comparing the recall in coding regions based on both dideoxy sequencing sets, we observe a significant difference in recall. Nevertheless, conclusions can only be made with caution because all estimates for the dideoxy set from Caicedo et al. (2007) are based on small numbers. Summarising, we cannot rule out the possibility that the TIGR dideoxy sequencing data is of lower quality than the polymorphism data set used in the *Arabidopsis* project.

Table 4.3: The performance of the MB method based on the data set provided from Caicedo et al. (2007). Recall and FDR are given in %.

Sequence Type	MB SNP calls	Recall	FDR
Coding	31	23.5	12.9
UTR	7	26.1	14.3
Intron	58	25.1	13.8
Intergenic	29	44.0	24.1
Total	125	27.1	16.0

4.4.3 Check of Paralogous Regions

In a further step, we addressed the question whether some fragments were amplified from a wrong location due to repetitiveness of the genome. As we did the k-mer analysis limited to the positions from the same amplicon pool, we did not have repetitive k-mer information across the whole genome. Having realised that the PCR primer had not been checked for uniqueness, we took into consideration that our dideoxy sequencing set contained ambiguous fragments. The sequenced fragments were aligned to the reference Nipponbare by Gunnar Rätsch to detect non-unique fragments allowing a range of different identity criteria. Table 4.4 shows the results for different BLAT searches. It is obvious that excluding paralogous frag-

ments improves the SNP calling performance. Recall can be more drastically increased than decreasing the FDR. Surprisingly, the FDR increases after excluding too many fragments.

Table 4.4: The number of SNP positions predicted by the MB method when fragments according to different BLAT criteria are excluded is shown. Recall and FDR are given in %.

BLAT Criteria	Excluded Fragments	MB SNP Calls	Recall	FDR
All	0	1,385	12.2	16.0
Length difference < 20 nt	96	1,318	12.5	15.5
90% identity	173	1,265	13.1	14.6
75% identity	262	1,222	13.3	14.6
50% identity	343	1,191	13.5	14.4
25% identity	459	1,136	13.9	14.7
More than one hit	802	1,010	14.5	15.2

4.4.4 Performance Dependency on Sequencing Quality Scores

Finally, we thought about filtering on the label level, i.e. excluding examples from the dideoxy sequencing set with respect to their sequencing quality score. Analogous to the experiment in Section 4.3.1, recall and FDR were determined discarding positions with sequence quality score below a varying threshold. In addition, we visualised the sequence quality score distribution of all predicted SNP positions compared to the distribution of all labelled SNP positions. Recall was calculated by adapting the total number of observed positives as we did not filter on the prediction, but on the label level. The results are shown in Figure 4.4. We observe that the recall increases from $\sim$12% to $\sim$15% where the steepest increase and thus the hightest gain is achieved for thresholds up to quality scores of 48. The same holds vice versa for the FDR which can be reduced to a level of 10% by applying a stringent quality score threshold. By filtering sequenced positions with a quality score less than 48, positions for which the base call was derived from only the forward or reverse read are discarded in large part.

4.4.5 Conclusion

From the assessment of the quality of our set of dideoxy sequences, we now have an idea how to incorporate this knowledge in evaluation of our SNP prediction algorithm. In general, a performance variation on a per strain basis is expected due to the fact that a differing amount of labelled data per strain is available for training. Furthermore, we showed that repetitive 25-mers increase the fraction of false positives by introducing a disturbing signal. We observed that the poor performance of SNP calling using the rice resequencing data is caused by diverse factors both on the level of predictions and the labelled data. It turned out that the quality of the set of dideoxy sequences can be improved by filtering with respect to the sequence quality score. We also realised that repetitiveness of the sequenced fragments reduces the reliability of SNP labels. Moreover, performance can be increased by discarding positions with low

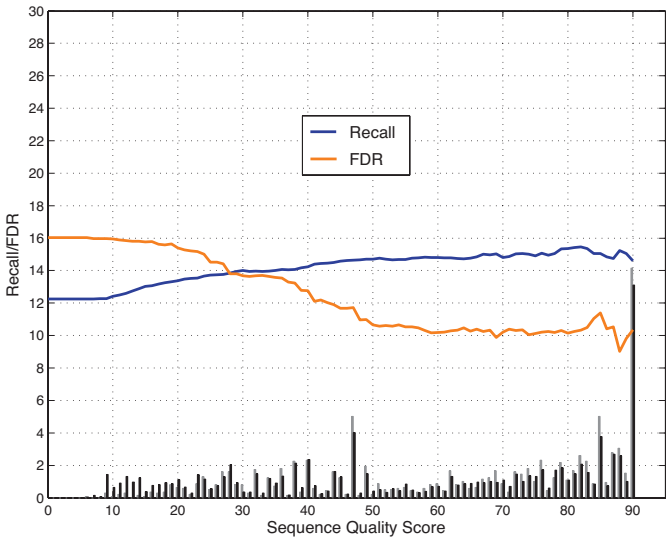


Figure 4.4: MB performance dependency on sequencing quality scores. The distribution of the sequence quality score at all predicted positions and at all labelled SNP positions are given in dark and light grey, respectively. Recall (blue) and FDR (orange) are shown as a function of the sequence quality score cut-off.

hybridisation quality score. Repetitive positions and those with low sequencing quality were used for training, but were excluded from performance evaluation.

However, the FDR is considerably higher compared to the FDR achieved for predictions by the same SNP calling method in the *Arabidopsis* resequencing project. Potential reasons for the worse performance remained largely unclear, but are probably attributable to experimental design and difference in experimental procedure between the rice and the *Arabidopsis* resequencing project.

5 SNP Calling

In this chapter, we address how to exploit the resequencing data in order to identify SNPs in all 19 strains, excluding the reference strain Nipponbare. We first introduce the basic concepts of the Support Vector Machine (SVM) classifier and then describe our approach to SNP calling. Thereafter, we present the results and give a performance analysis of our machine learning approach in comparison to the model based approach by Perlegen Sciences. The chapter is concluded with a discussion and an outlook to future analyses.

Using hybridisation data of resequencing experiments, two signals can be examined for the discovery of SNP sites in general. A base call which does not correspond to the base of the reference sequence at a particular site is inferred by both a decreased intensity of the reference probe (loss-of-signal) and an increased intensity of a non-reference probe (gain-of-signal) (see Hacia, 1999). As a SNP at the centre position of a probe introduces mismatches for adjacent probes, loss-of signal of surrounding probes can be exploited in SNP calling approaches as well.

Computational methods which are based on the interpretation of hybridisation data have been developed for identification of polymorphic sites. A simple approach using statistical models and assigning quality scores to each site was published by Cutler et al. (2001) for the application to Affymetrix Variation Detection Arrays. In Section 4.3, we described another model-based, more advanced method by Perlegen Sciences. Another approach by Zhan and Kulp (2005) incorporates physical properties of hybridisation between two oligonucleotides to build a base-calling method. In the next section, we present an alternative method for SNP calling from resequencing data using machine learning as previously used in the *Arabidopsis* resequencing project (see Clark et al., 2007).

5.1 A Two-Layered Machine Learning Method for SNP Calling

The goal of machine learning is to automatically detect relations between samples of a certain data set. In terms of the SNP identification problem the task is to learn the characteristics which discriminate SNP and non-SNP positions using hybridisation measurements. Such a learning task belongs to the so-called binary or two-class classification problems. In our case, SNP positions form one class, intuitively labelled as the positive class. A second, negative class is defined by all positions which are invariant. Now, the goal is to find a function which separates these two classes best. To approach such problems, one family of methods are Support Vector Machines.

5.1.1 Support Vector Machines (SVMs)

In the following, we rather focus on the basic ideas and concepts of Support Vector Machines than giving a detailed introduction to the method. More details on SVMs and kernel methods can be found in Vapnik (1998, Chapter 10), Burges (1998) and Schölkopf and Smola (2002).

Support Vector Machines are a set of methods which achieve the minimisation of the classification error by maximising a so-called margin. The main idea of SVMs is to find the hyperplane which separates the data with the largest margin. For separable problems, the margin is defined as the distance between the closest positive example to the hyperplane and the closest negative example to the hyperplane .
From a geometric point of view such a separating hyperplane can be described as $\langle \mathbf{w}, \mathbf{x}_i \rangle + b = 0$ if there is a unit vector $\mathbf{w}$, which is the normal to the hyperplane, and a constant b such that the inequality $y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 0$ holds for all $i = 1, \ldots, n$ where $\mathbf{x}_i \in \mathcal{X}$ are the input examples with corresponding class label $y_i \in \{-1, +1\}$.
Mathematically spoken, the SVM algorithm aims to find a function

$$f(\mathbf{x}) = \text{sign}(\langle \mathbf{w}, \mathbf{x} \rangle + b),$$

where $\mathbf{w}$ and b are found by solving the convex optimisation problem which determines the maximum margin hyperplane:

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + \gamma \sum_i^n \xi_i \\ \text{subject to} \quad & y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1 - \xi_i \qquad \text{and} \qquad \xi_i \geq 0 \\ & \text{for all } i = 1, \ldots, n \end{aligned}$$

The definition above allows a certain amount of misclassification by introducing so-called slack variables $\xi_i,\ \xi_i \geq 0, i = 1, \ldots, n$. The penalty for using slack variables is represented by the hyperparameter γ as an error weight on the outliers. Since the margin constraint is softened by using slack variables, this SVM formulation is also known as a soft-margin SVM. One approach for dealing with such kinds of constrained optimisations problems is to introduce Lagrange multipliers $\alpha_i \geq 0$ and a Lagrangian

$$\mathcal{L}(\mathbf{w}, b, \alpha) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^n \alpha_i (y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1).$$

After some mathematical transformations and the elimination of $\mathbf{w}$ and b one arrives at the so-called dual optimisation problem:

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle \qquad (5.1) \\ \text{subject to} \quad & \sum_{i=1}^n \alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq \gamma \text{ for all } i = 1, \ldots, n \end{aligned}$$

All data points which are not necessary to construct the optimal separating hyperplane have Lagrange multipliers equal to zero. All other points are called support vectors, i.e. their

corresponding Lagrange multipliers are greater than zero. The support vectors lie on one of the margin hyperplanes or inside the margin as their α_i give the solution of the maximal margin.

Until now, we have discussed SVMs as linear classifiers, i.e. the decision function was a linear function of the data. The question we want to address now is what we can do when dealing with nonlinear class boundaries.

The general idea is to map data from an input space to a high dimensional feature space $\mathcal{H}$ and separate the images of the input data in this space by a linear decision function:

$$\begin{aligned} \phi : \mathcal{X} &\longrightarrow \mathcal{H} \\ \mathbf{x} &\longmapsto \phi(\mathbf{x}) \end{aligned}$$

where $\mathcal{X}$ is the input set. Note that although the decision boundary is linear in $\mathcal{H}$, it can be highly nonlinear in $\mathcal{X}$.

Since the SVM objective 5.1 is formulated in terms of dot products between the examples, we do not have to compute $\phi(\mathbf{x}_i)$ for all $\mathbf{x}_i \in \mathcal{X}$ explicitly. It suffices to use a function which computes the dot product in the feature space. Such a function is called a kernel and is defined as a function k that satisfies

$$k(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle$$

$$\forall\ \mathbf{x}, \mathbf{y} \in \mathcal{X}.$$

Computing only dot products in the feature space and expressing the dot product in terms of a kernel k is also known as the *kernel trick*. Like the dot product is a similarity measure in the input space, the kernel is a measure of similarity in the feature space. The benefit of using a kernel function is that the mapping ϕ does not have to be known explicitly to solve the learning problem in the potentially very high dimensional feature space.

The formula 5.1 can now be rewritten in terms of a kernel function k:

$$\begin{aligned} \underset{\alpha}{\text{maximise}} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) \\ \text{subject to} \quad & \sum_{i=1}^{n} \alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq \gamma \text{ for } i = 1, \ldots, n \end{aligned}$$

A variety of kernels have been developed and applied to numerous different classification problems (see e.g. Schölkopf and Smola, 2002). One of the kernels which has been extensively used and which we also employed for SNP identification is the radial basis function (RBF) or Gaussian kernel

$$k(\mathbf{x}, \mathbf{y}) = \exp\left(\frac{-||\mathbf{x}-\mathbf{y}||^2}{2\sigma^2}\right).$$

The paramater σ of RBF kernel controls the width of the Gaussian function. As a variety of different functions can be generated by tuning σ, this kernel is applicable to a broad set of problems.

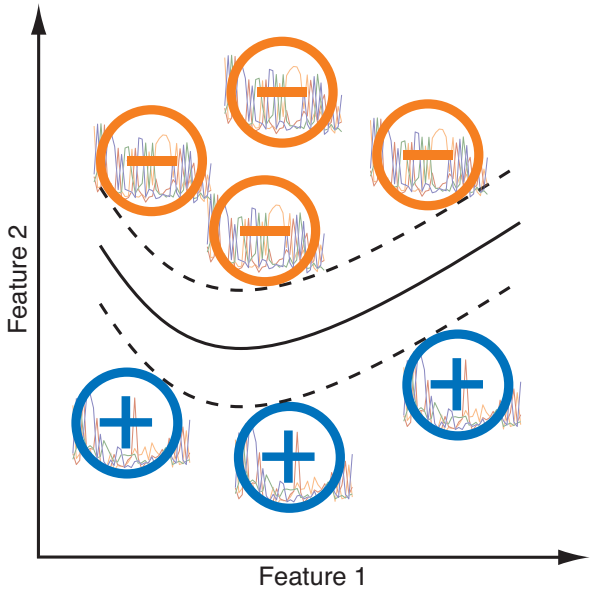


Figure 5.1: The picture outlines the concept of SVMs applied to SNP classification. The problem is simplified to a two-dimensional feature space. The feature vector corresponding to SNP positions are represented by blue circles. Negative examples are coloured in orange. The separating decision boundary is shown as the black solid line. The support vectors are those data points lying on the margin boarders (drawn as dashed lines).

5.1.2 Design of the Two-layered SVM Method

We applied a two-layered approach based on SVMs to predict SNPs from the hybridisation data similar to the approach described in the paper by Clark et al. (2007) for the *Arabidopsis* resequencing project. In a first step, SVMs were strain-independently trained using information comprising the array data, sequence characteristics and repetitiveness of the genome using the results of the annotation of repetitive 25-mers. Since the hybridisation had been normalised beforehand, we were able to train machines across all strains instead of using a separate machine for each strain as done for *Arabidopsis*. Subsequently, knowledge gained from layer 1 predictions was added to train a second layer of SVMs. According to biological observations, singleton SNPs occur rarely, i.e. it is likely that a SNP in a particular strain is identified when having observed this SNP in at least one other strain before. The second layer took advantage of this observation by including outputs of the first layer. We applied the trained SVMs of the second layer for final genome-wide predictions.

Each layer was divided into several subtasks. First, we generated input vectors for positions which passed certain filter criteria. After model selection in a cross-validation procedure, SVMs were retrained with the optimal model parameters to obtain predictions for the filtered positions across all strains. Finally, each prediction was transformed to a confidence value reflecting the posterior likehood of a true SNP prediction.

In order to train the SVMs and to evaluate the performance of the classifier, the data set of known polymorphic sites described in Chapter 4 was utilised.

The next two sections describe the applied methods in more detail based on the definitions given in Clark et al. (2007).

5.1.3 Layer 1 SVM

5.1.3.1 Filter Criteria

By applying a filter prior to the training step, we intended to increase the fraction of true polymorphic sites, leading to a more balanced set of training examples which is less challenging for accurate discrimination. Indeed, a large fraction of the non-polymorphic sites were thereby discarded. This resulted in a significantly smaller data set being beneficial to reduce computational time in both training and prediction.

For this purpose, we excluded positions which were identical to the reference with high probability. Furthermore, positions were discarded if the corresponding array data gave inconsistent information on the called base.

The first criterion was met for a given position p in the target strain t if the raw base call $B_t^+(p)$ and $B_t^-(p)$ of the forward and reverse strand were identical, but different from the base $RS(p)$ of the reference sequence. Secondly, the reference raw base calls of both strands $B_{ref}^+(p)$ and $B_{ref}^-(p)$ had to be consistent to each other and to $RS(p)$. To discard regions of amplicon failures, we also excluded positions with hybridisation quality scores $\overline{QS}_t(p)$ less than or equal to 5, averaged over a 100 bp window.

Taking these criteria together, we can describe all positions P which passed the filter as the set:

$$\begin{aligned} P &= \{p_t | t \text{ a given target strain}\} \\ \text{where } p_t &= \{p | B_t^+(p) = B_t^-(p) \neq RS(p) \wedge B_{ref}^+(p) = B_{ref}^-(p) = RS(p) \\ &\quad \wedge \overline{QS}_t(p) > 5\} \end{aligned}$$

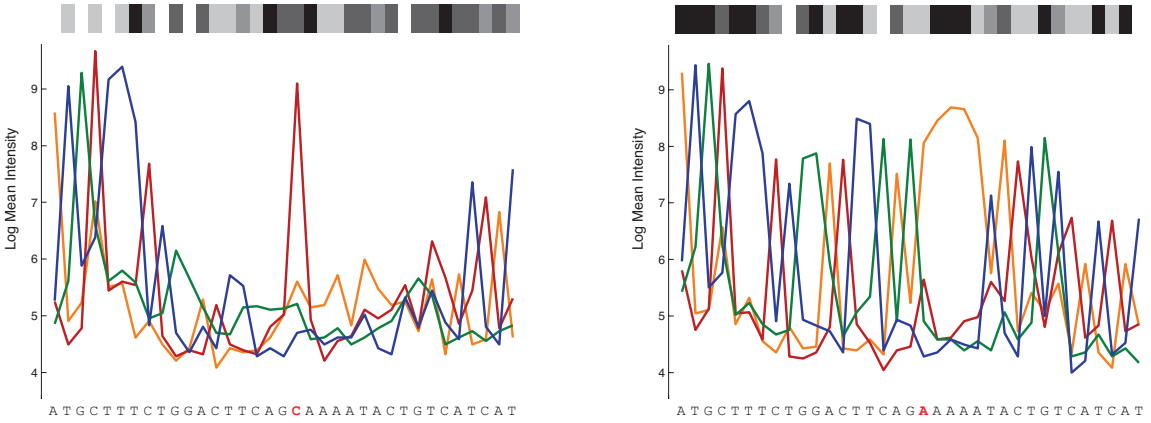
5.1.3.2 Input Generation

To be able to train SVMs on the given data set, we had to generate input vector $\mathbf{x}_p^1$ for each position p which had passed the filter in layer 1. Each $\mathbf{x}_p^1$ includes measurements at this position and neighbouring positions within ± 4 bp around p: Both maximal intensities I_{max} of each quartet respectively and averages of the non-maximal intensities I_{sec} of each quartet respectively at each position in the 9 bp window were included. As the machine learning method was more accurate for SNPs separated by 7 to 30 bp according to observations for SNP detection in *Arabidopsis thaliana*, we chose a window size of 9 bp to have predictions complementing the SNP calls made by the model based method.

Moreover, we added ratios of the maximal intensities at p and its neighbouring positions (Q_1) and ratios of the maximal intensities at p of the target and the reference strain (Q_2). The usage of these quotients as input features was motivated by the shape of a typical SNP signal reflected in the intensity pattern (see Figure 5.2). $\mathbf{x}_p^1$ also contains sequence characteristics such as mismatches M between raw base calls and the reference sequence within the window, the reference base RS, frequencies f of each base (`A`, `C`, `G`, `T`) within the 25-mer and the sequence entropy H of the 25-mer. Furthermore, we used the results from the 25-mer analysis to include occurrence counts k of repetitive 25-mers at p.

As we used strain-independent training, we included information on the strain of which the position was taken, denoted as str. Table 5.1 describes all inputs in more detail.

The input vectors were normalised on the training set to mean 0 and standard deviation 1 per input dimension. Additionally, a normalisation per input example was applied using $\frac{\mathbf{x}_p^1}{||\mathbf{x}_p^1||}$ for all positions p. The normalised input vectors were then used to train SVMs with a RBF kernel (see Section 5.1.1 on SVMs) using the SHOGUN toolbox,[1] developed by Sonnenburg et al. (2006), which allows a fast and efficient training.



(a) Minghui 63 (b) Nipponbare

Figure 5.2: Intensity measurements are shown for the *indica* strain Minghui 63 and the reference Nipponbare from position 3,737,196 to 3,737,230 on chromosome 1. Orange denotes the mean intensities averaged over both strands corresponding to the base `A`, red to `C`, green to `G` and blue to `T`. The bar on top encodes the mean quality scores at the positions where a darker colour corresponds to a higher mean quality score. (a) The picture shows a prototypical SNP signature around position 3,737,214 for Minghui 63. The intensities around the SNP positions are reduced due to mismatches in the neighbouring probes. (b) In contrast to that, the intensity patterns for the reference are visualised.

5.1.3.3 Model Selection in a Cross-validation Procedure

Cross-validation is a procedure in which a given data set is partitioned into subsets to have disjoint sets for training and performance evaluation in permutated order, also important for model selection. We randomly split our data set of labelled positions into five equally sized disjoint subsets $S_1, \ldots, S_5$ with respect to an uniform distribution of positives per strain. Training and model selection was performed on five different folds in a nested cross-validation scheme: At each of the five iterations a different subset S_i served as test set T_i. The set which was used for model selection was then defined as $X_i = \{S_j | j \neq i\}$. Thus, the set X_i at iteration step i consisted of 80% of all labelled positions, whereas the remaining positions belonged to T_i. For parameter tuning, each set X_i, $i = 1, \ldots, 5$ was in turn partitioned into five subsets S_{ij}, $j = 1, \ldots, 5$, four of them served as training set X_{ij} at each iteration step. The prediction and evaluation was done on the omitted subset for each model k, i.e. for each combination of the model parameters. The parameters to be tuned comprised the width σ of the RBF kernel ($\sigma = [10^{-1}, 10^{-0.5}, 10^0, 10^{0.5}, 10^1]$) and the penalty for using slack variables ($\gamma = [10^{-2}, 10^{-1}, 10^0, 10^1]$). Thus, 20 models were tested on each subset S_{ij}.

[1] `http://www.fml.tuebingen.mpg.de/raetsch/projects/shogun`

To find the best model, we applied a measurement based on the idea of the area under the curve (AUC) of the receiver operating characteristic (ROC). ROC curves and their AUCs are commonly used for the performance assessment of a binary classifier. In a ROC curve, the false positive rate (sensitivity) is plotted against the true positive rate ($1 -$ specificity). The higher the AUC value is, the more accurate the classifier. In our case, we only wanted to maximise the area limited to a maximal ratio of true positives to false positives, in other words, the AUC for a set of classifiers with a limited FDR, as we were interested in a classifier optimal up to this FDR. We therefore determined the number of true positives as a function of the number of false positives. Then, the area a_{ikj} between this curve and the line showing one false positive at five true positives was calculated corresponding to FDRs below $\frac{1}{6}$.[2] The optimal model for each split was determined by the model k for which the average of the areas a_{ikj}, $j = 1, \ldots, 5$ was maximal. Thereafter, a SVM on the whole set X_i was trained with its best model parameters and predictions were calculated for the hold-out set T_i.

5.1.3.4 Prediction and Output Transformation

The prediction for each position which passed the filter was computed by the SVM of layer 1 for which this position was not used in training or parameter selection. For all other positions, we could use any of the trained SVM. Thus, we randomly chose one of the five layer 1 SVMs to predict their labels.

As we used outputs from five different SVMs, the predictions were not directly comparable. In order to be able to employ these predictions as an input to the layer 2 SVMs and for further analyses, each prediction was transformed into a posterior probability for being a true positive. For this purpose, the conditional likelihood of the true label being positive for a given output value was estimated.

The aim was to learn a piecewise linear function on the corresponding test set. The SVM predictions were divided into 40 quantiles of which each was represented by a supporting point $x(q)$, $q = 1, \ldots, 40$ to ensure a good estimation of the piecewise linear function. The probability $y(q)$ of being a true positive was estimated as

$$y(q) = \frac{n_{TP}(q)}{n(q)}$$

where $n_{TP}(q)$ is the number of true positives, i.e. the number of known SNP positions, with prediction values V in the range $x(q) \leq V \leq x(q+1)$ and $n(q)$ the number of all labelled positions with prediction values in that range. Analogous estimations were made for definition of a cumulative y_c by omitting the upper bound. To obtain smooth and monotonically increasing estimates, a technique described in Sonnenburg et al. (2007) was used.

Each prediction value V was transformed into a confidence c by:

$$c = \begin{cases} y(1) & V \leq x(1) \\ \frac{y(q+1)\cdot(V-x(q))+y(q)\cdot(x(q+1)-V)}{x(q+1)-x(q)} & x(q) \leq V \leq x(q+1) \\ y(40) & V \geq x(40) \end{cases}$$

The transformation function for the cumulative confidence C worked in a analogous manner.

[2] In Clark et al. (2007), the threshold was set to 0.05.

5.1.4 Layer 2 SVM

5.1.4.1 Filter Criteria

In the second step, we used one additional condition by exploiting information from layer 1 predictions across all strains. Only positions which had confidence values c_t greater than some threshold th_t in at least one other strain t were used for training the SVMs of layer 2. This threshold was determined on all test sets per strain by taking the confidence value th_t above which n_t examples had transformed prediction values, where n_t is the sum of all positively labelled positions in t which passed the filter. We further used a relaxed filter for layer 2 to take positions into account which were likely to be polymorphic when observed in at least one strain, but which did not have raw base calls identical on both strands in the strain and the reference, respectively. Furthermore, the raw base calls of at least one strand had to be different to the reference sequence. For the reference raw base calls, an inconsistency to the reference sequence in one of the strands was allowed. We again discarded position with mean hybridisation quality score $\overline{QS}_t(p)$ less than or equal to 5.

All positions P which passed filter 2 are described as:

$$
\begin{aligned}
P &= \{p_t | t \text{ a given strain}\} \\
\text{where } p_t &= \{p | \bigvee_{s=1}^{19} [[c_s(p) > th_s]] \wedge (B_t^+(p) \neq RS(p) \vee B_t^-(p) \neq RS(p)) \\
&\wedge (B_{ref}^+(p) = RS(p) \vee B_{ref}^-(p) = RS(p)) \wedge \overline{QS}_t(p) > 5\}
\end{aligned}
$$

where [[ ]] denotes the indicator function.

5.1.4.2 From Input Generation to Predictions

The input vector $\mathbf{x}_p^1$ from layer 1 was extended to the input vector $\mathbf{x}_p^2$ for layer 2 by a binary vector b having ones at positions for which the corresponding confidence values were above the threshold th_t. In addition to that, the confidence values of all strains were included. To be able to connect a position to its target strain t, we encoded the strain information by vector of length $19^2 = 361$ with b at the 19 positions corresponding to strain t and zeros elsewhere. Confidence values were encoded in same way. The strain information *str* of a position was omitted. As an additional feature, we included knowledge on variation between the genome sequence of the *ssp. japonica* strain Nipponbare and the *ssp. indica* strain 93-11 (*ind*), facilitating the detection of SNPs at those observed polymorphic positions. We aligned the two sequences using MUMMER,[3] a suffix-tree algorithm for large-scale genome alignments by Kurtz et al. (2004), and parsed SNPs from the alignments with more than 90% identity within a 61 bp window. See Table 5.1 for more details on the layer 2 inputs.

After normalisation of the input vectors, we applied the same model selection procedure as for layer 1. Predictions were made for all positions which passed filter 2 by exactly one layer 2 SVM.

[3] `http://mummer.sourceforce.net`

5.1.5 Predictions for All Arrayed Positions

After having trained the five layer 1 SVMs and five layer 2 SVMs based on the data set of known SNPs, we then applied these trained machines to make predictions for the tiled regions of the whole genome including unlabelled positions which were interrogated on the hybridisation resequencing arrays. As theses positions had not been employed either for training and evaluation, any of the five SVMs could be used for prediction.

For all of the 19 non-reference strains, a layer 1 SVM was chosen at random to make a prediction for each unlabelled position which passed filter 1. We also made predictions for positions across all strains which met the filter criteria in at least one other strain. Afterwards, the outputs were transformed into confidence values applying the transformation function specific to the layer 1 machine used. This way, the predictions could be exploited for the second prediction layer.

For all unlabelled positions which were predicted by layer 1 SVMs and passed the second filter, predictions were made by all five layer 2 SVMs. The SVM outputs were again transformed into confidences with the corresponding piecewise linear function. The resulting five values for each position were averaged to assign a final probability of being a SNP position.

5.2 Results

We exploited the data set of known polymorphism described in Chapter 4 containing 1,755,395 positions with 9,499 SNPs (∼5 SNPs per kb) in training and evaluation of the SVMs. After applying the filter of layer 1, the size of the data set was reduced to 20,133 positions with 1,946 positives. Thus, 98.85% of all positions in the dideoxy data set were discarded, keeping only ∼20% of all SNP positions. By filtering the ratio of positive examples, i.e. true SNP positions to all examples is increased from ∼1:185 to ∼1:10.

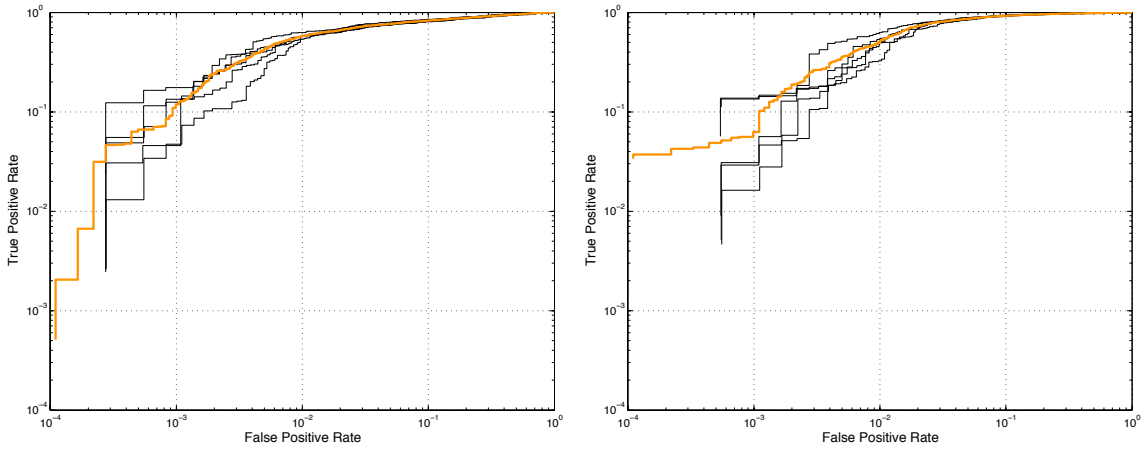
After training and predicting with the layer 1 machines, SNP calls for 17,861 unique positions across all 19 strains were available. The filter for layer 2 further reduced the number of remaining positions to 11,194, while retaining the number of true SNPs (2,093) in a magnitude comparable to the number of SNPs which passed filter 1. Thus, the ratio of positive to all examples was increased to ∼1:5. Note that the recall is limited by the number of true SNP positions which passed filter 2, i.e. at most 22% of true SNPs could be predicted by the ML method.

Genome-wide predictions by our machine learning based method resulted in the detection of SNPs at 1,366,243 positions. All predictions made by the ML method with cumulative confidence greater than or equal to 0.855 were taken for final MB SNP calls. The combination of the SNP predictions made with the MB and ML method formed a set of 1,930,400 SNP sites of which 421,965 were nonredundant. 763,211 (39.5%) of these positions were predicted by both methods, 564,157 (29.2%) by the MB method only and 603,032 (31.2%) by the ML method only. Table A.7 shows the number of SNP calls for both methods for each strain. More SNPs were called by the ML method for all strains except for the *temperate japonica* strains, i.e. the SNP prediction was more difficult for the ML method for these three strains. Only two SNPs were predicted for Tainung 67 by the ML method, reflecting lower cumulative confidence values for this strain.

5.2.1 Performance Analysis

5.2.1.1 General Performance

The performance of the SVM classifiers was evaluated against the dideoxy data set. For this purpose, we first determined the area under the ROC curve (see Section 5.1.3.3) for all five SVMs of both layers. The resulting curves are given in Figure 5.3. The AUC value averaged over the five layer 1 SVMs scored to 92.6%. This would suggest a particularly good classifier if the data set was balanced. The very low standard deviation reflected that the SVMs of the five different splits performed similarly well. The performance analysis of the layer 2 SVMs resulted in a mean AUC of 96.3%, showing an even better performance.

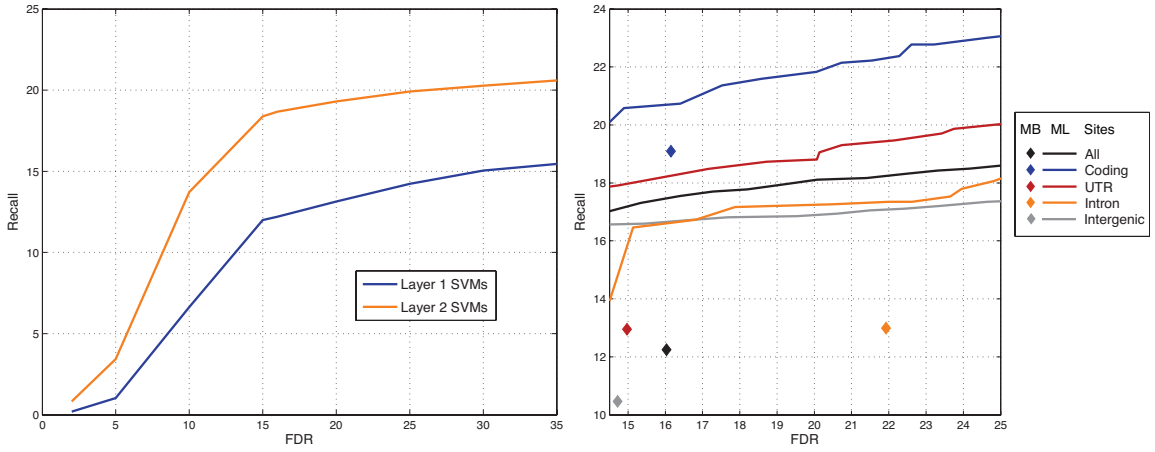


(a) Layer 1 SVMs (b) Layer 2 SVMs

Figure 5.3: Both figures show the ROC curves for each of the five splits (black) and all predictions together (orange). (a) The five layer 1 SVMs performed with an AUC value of 92.6% on average with standard deviation ±0.010. (b) An AUC of 96.3% on average with standard deviation ±0.004 was achieved by the five layer 2 SVMs.

Receiver operating characteristics are insensitive to class balance as they show the true positive rate as a function of the false positive rate. We therefore used as a second performance measure similar to precision recall curves (PRCs) where the recall is plotted versus the precision. We examined the performance by determining recall as a function of different FDRs for each layer separately. Therefore, we exploited the fact that the cumulative confidence value C corresponds to the precision and thus defines the FDR by $1 - C$. That means that e.g. for all SVM predictions with a C value greater than 0.85, 15 false positives are expected in a set of 100 predictions. Hence, we could calculate the recall defining predictions with a C value above a threshold as positives, those below the threshold as negatives. Figure 5.4(a) shows the resulting curves for layer 1 SVMs and layer 2 SVMs. We observed a rapid increase of the recall values in the FDR range from 5% to 15%, i.e. the number of true positives increases faster than the number of false positives. Above a FDR value of 15%, the increase in recall rate was smaller reflecting that a higher number of true positives found also implicates more false predictions. Moreover, the figure shows the benefit of the two layered approach. The layer 2 SVMs outperformed the layer 1 SVMs: Using the knowledge gained by layer 1

SVMs predictions, recall could be improved by about six percentage points on average. However, due to the design of the layer 2 SVMS, the predictions were potentially biased towards polymorphic sites which were observed in the *indica* strain 93-11 and towards SNPs which occurred in more than one strain.



(a) (b)

Figure 5.4: (a) FDR-dependent recall is shown for layer 1 SVMs (blue) in contrast to the performance of the layer 2 SVMs (orange). (b) The two SNP detection methods MB and ML are compared with respect to their recall and FDR. Overall FDR-dependent recall is coloured in black. Performance with respect to the different sequence types coding (blue), UTR (red), intron (orange) and intergenic (grey) is illustrated as well.

5.2.1.2 Performance Analysis Using Filters

In addition, the performance was calculated applying several filter criteria motivated by the observations described in Chapter 4. The results are given in Table A.6. With a specificity threshold of $C = 0.84$, a recall of 17.4% was achieved at a FDR of 16.0%, considering all SNP positions predicted by the ML method at once. Performance varied between the five different subspecies to a larger extent than within subspecies. Best recall rates were observed for the *indica* and *aus* subspecies. Predictions for the *indica* strain SHZ2 were most accurate with a recall of 17.2% at a FDR of 11.3%. Similar performance was achieved for the two *aus* strains Rayada and FR 13A and for the *indica* strain Minghui 63. The better performance for the strains of these two subspecies can be explained by the layer 2 SVM design biased for *indica* strains since SNPs from the strain 93-11 were included as a feature and the labelled data set was overbalanced for *indica* and *aus* strains. For the three *temperate japonica* strains closely related to Nipponbare and for the strain Dom Sufid, we noticed the worst performance. Note that for these strains the performance measures were not meaningful as only a very few SNP calls were made.

In a next step, the influence of repetitive positions identified by the 25-mer analysis on the ML SNP predictions was investigated. For this purpose, we excluded either all positions which were annotated as repetitive with respect to all matching criteria, positions with exact 25-mer matches or positions with dominating repetitive 25-mer matches. The exclusions of

repetitive 25-mer positions had no significant effect on the performance given that only a few positions ($< 3.2\%$) were discarded by these filters. This result was expected as information on repetitive sites was included in the input vectors of the SVMs and could thus be employed in the training step for discrimination between true SNPs and spurious SNP signals caused by cross-hybridisation.

We also filtered by hybridisation quality score including only those positions for which the mean hybridisation quality score was above a threshold of 8. Although Figure A.9 would propose a conservative threshold of 2, we decided to use the same threshold as for the MB performance analysis to obtain comparable results. We observed a reduced FDR of 12.7%, decreasing the recall rate by four percentage points at the same time.

In another experiment, we studied the classifier performance filtering on the level of the labelled data set of dideoxy sequences (see Figure A.10). After we had excluded the positions with sequencing quality score less than 48, the recall was increased to 21.7% at a FDR of 9.7%. A smaller performance improvement was observed when discarding labelled positions from fragments which were more than 90% identical to at least one other region of the genome. Furthermore, we only considered positions for which the labels were identical in at least two strains to measure the performance. In this way, we disregarded singleton SNPs to obtain more reliable dideoxy labels. Filtering by this criteria resulted in an unchanged recall at a reduced FDR of 15.1%.

5.2.2 Comparison to the Model Based Method

In order to evaluate the performance of the machine learning method, we further compared our results with the SNP predictions made by the MB method by Perlegen Sciences (see Section 4.3 for details on the method). The exclusion of repetitive 25-mer positions in the evaluation of the MB method had a similar effect than for the ML method. Filtering singleton SNPs improved the performance to only a small extent for both methods, supporting the conclusion we arrived at in Chapter 4 that sample contamination could be excluded as a source for false SNP labels. Comparing the performance on a per strain basis, we observed the largest differences for strains of the *indica* subspecies. Here, the ML SNP calls were more accurate as we used *indica* SNPs as a feature for layer 2 SVMs. For both methods, only a very small number of SNPs were called for Dom Sufid, reflecting a general poor performance for this strain.

In general, with the machine learning method, more SNPs could be recovered than with the MB method (17.4% versus 12.2% recall at the same FDR of 16%). We also determined the recall as a function of the FDR on a sequence type basis. Figure 5.4(b) illustrates the results. Similar to the results of the *Arabidopsis* resequencing project, the performance of the ML method was inferior to the MB method for UTRs, introns and intergenic regions with a high difference in recall rates up to seven percentage points. In contrast to the observations for *Arabidopsis* SNP calls, the ML method also clearly achieved better results than the MB method for coding regions with a smaller difference in recall of three percentage points. Moreover, the performance of the SNP detectors in coding regions was significantly higher than for non-coding sequences. We explain this by the observation that coding sequences are favourable for array-based SNP detection due their different sequence composition and low polymorphism levels.

Based on the performance results allowing different filter criteria, we decided to exclude all sites with repetitive 25-mer matches and positions with sequencing quality score less than 48 for final performance evaluation. For the evaluation of the MB method, we further excluded positions with hybridisation quality score less than 8. Table 5.2 shows the resulting recall rates and FDRs for both methods. By applying this filter, FDRs could be reduced to less than 10% for almost all strains. Altogether, 20.9% of polymorphic sites were recovered by the ML method at a FDR of 8.3%, clearly superior to the performance of the MB method (13.9% recall at a FDR of 7.8%).

We also studied the performance taking MB and ML SNP calls together. On the combined set of MB/ML SNP calls more SNPs were recovered as expected (23.8%) at a slightly worse FDR of 10.3%. The FDR was reduced to 2.9% on the set considering SNPs predicted by both methods. Thus, these SNPs calls are the most reliable predictions. Moreover, 9.9% of SNP sites were recovered by the ML method only at an increased FDR of 13.7%, considerably superior to the performance for the predictions made with the MB method only with a very low recall of 3.2% and a high FDR of 21.1%.

5.3 Conclusions

Summarising, we presented a machine learning approach based on Support Vector Machines to identify polymorphic sites in the non-repetitive portion of the genomes of 19 diverged rice strains. In comparison to the model based method by Perlegen Sciences, we designed a classifier with higher recall at the same FDR. Both methods provided complementary predictions of polymorphic sites having 39.1% of the predictions in common. 29.1% (MB) and 31.5% (ML) of the polymorphic sites were predicted with exclusively either one of the methods. However, on the rice hybridisation data both methods did not perform as good as on the array data used in the *Arabidopsis* resequencing project.

Table 5.1: Properties of the input vectors $\mathbf{x}_p^1$ for layer 1 and $\mathbf{x}_p^2$ for layer 2 at a given position p are shown. $\Sigma = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ denotes the DNA alphabet, the strands are represented by $s \in \{+,-\}$, τ is either target t or reference strain ref. Unless defined otherwise, $\Delta p \in \{-4, \dots 4\}$.

	Formula	Description	Size
I_{max}	$I^p_{max}(\Delta p, \tau, s) = \max_{\sigma\in\Sigma} \log\left(I_{\tau,s}(p+\Delta p, \sigma)\right)$	Maximal intensities for target and reference strain, strand-wise in a 9 bp window	36
I_{sec}	$I^p_{sec}(\Delta p, \tau, s) = \text{mean}_{\sigma\neq\sigma_{max}} \log\left(I_{\tau,s}(p+\Delta p, \sigma)\right)$ where $\sigma_{max} = \arg\max_{\sigma\in\Sigma} \log\left(I_{\tau,s}(p+\Delta p, \sigma)\right)$	Average of the non-maximal intensities for target and reference strain, strand-wise in a 9 bp window	36
Q_1	$Q^p_1(\Delta p, \tau, s) = \log\left(\frac{I^p_{max}(\Delta p,\tau,s)}{I^p_{max}(0,\tau,s)}\right)$ where $\Delta p \in \{-4, \dots, -1, 1, \dots 4\}$	Quotients of the maximal intensities at the neighbouring positions and p for target and reference strain, strand-wise in a 9 bp window	32
Q_2	$Q^p_2(\Delta p, s) = \log\left(\frac{I^p_{max}(\Delta p,t,s)}{I^p_{max}(\Delta p,ref,s)}\right)$	Quotients of the maximal intensities at p for target and reference strain, strand-wise in a 9 bp window	18
M	$M^p(\Delta p, \tau, s) = \delta(B_{\tau,s}(p+\Delta p), RS(p))$ where $\delta(i,j) = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$	Mismatches between the raw base calls of the target/reference strain and the reference sequence, strand-wise in a 9 bp window	36
RS	$RS^p(\sigma) = \delta(RS(p), \sigma)$	Binary vector indicating the presence of the base in the reference at p	4
f	$f^p(\sigma) = \sum_{\Delta=-13}^{\Delta=13} \delta(RS(p+\Delta), \sigma)$	Frequency of each base within the 25-mer	4
H	$H^p = -\sum_{\sigma\in\Sigma} f^p(\sigma)\log(f^p(\sigma))$	Sequence entropy of the 25-mer	1
k	$k^p(\Delta p, \sigma) = [k^p_{type}(\Delta p, \sigma), k^p_{dom\ type}(\Delta p)]$ where $type \in \{exact, inexact, short, bulged\}$	Occurrence counts of repetitive 25-mers in a 9 bp window	558
str	$str^p(\gamma) = \begin{cases} 10 & \gamma = str(p) \\ 0 & \gamma \neq str(p) \end{cases}$	Vector indicating the strain origin of p	19
	$\mathbf{x}^1 = [I_{max}, I_{sec}, Q_1, Q_2, M, RS, f, H, k, str]^T$		744
b	$b^p(t) = [[c^p(t) \geq th_t]]$	Binary vector indicating whether p passed th_t for strain t	361
c	$c^p(t)$	Confidences at p for all strains t	361
ind	$ind^p = \begin{cases} +1 & RS_{jap} \neq RS_{ind} \\ -1 & RS_{jap} = RS_{ind} \end{cases}$	Variation to the *ssp. indica* reference	2
	$\mathbf{x}^2 = [I_{max}, I_{sec}, Q_1, Q_2, M, RS, f, H, k, b, c, ind]^T$		1,449

Table 5.2: Shown is the number of MB and ML SNP calls after filtering out positions with repetitive 25-mer matches and sequencing quality score less than 48. In addition, MB SNP calls at positions with hybridisation quality score less than 8 were excluded. We grouped the strains with respect to their population affiliation. The two performance measures [recall, FDR] are given in % in squared brackets. N/A denotes FDRs which could not be assessed as no predictions were made.

Strain	MB SNP calls	ML SNP calls
LTH	4 [66.7, 00.0]	1 [16.7, 00.0]
M 202	4 [33.3, 00.0]	1 [08.3, 00.0]
Tainung 67	2 [66.7, 00.0]	0 [00.0, N/A]
Azucena	13 [09.8, 15.4]	28 [20.5, 17.9]
Cypress	24 [11.1, 12.5]	36 [16.4, 13.9]
Moroberekan	11 [06.3, 09.1]	20 [11.9, 05.0]
Dom Sufid	3 [04.7, 33.3]	0 [00.0, N/A]
Dular	55 [12.8, 07.3]	89 [20.2, 09.0]
FR 13A	64 [15.0, 07.8]	109 [24.1, 12.8]
N 22	78 [14.3, 06.4]	104 [18.1, 11.5]
Rayada	94 [17.4, 02.1]	146 [26.3, 04.8]
Aswina	46 [17.9, 06.5]	65 [25.4, 06.2]
IR 64	48 [14.6, 08.3]	77 [24.8, 02.6]
Minghui 63	91 [17.4, 05.5]	139 [26.3, 06.5]
Pokkali	62 [10.0, 09.7]	106 [17.1, 09.4]
SHZ2	75 [15.2, 12.0]	90 [19.4, 06.7]
Sadu Cho	38 [12.0, 07.9]	57 [18.5, 05.3]
Swarna	75 [12.8, 06.7]	119 [20.7, 05.0]
Zhenshan 97	60 [13.3, 13.3]	92 [19.9, 15.2]
Total	847 [13.9, 07.8]	1,360 [20.9, 08.3]

6 Conclusion and Future Work

In this thesis, I report my analysis of a rich set of array resequencing data, generated to reveal global polymorphism patterns in rice. Sophisticated computational approaches were implemented for efficient and accurate analysis of this data set.

Prior to the actual SNP analysis, a set of preprocessing steps were undertaken. We first examined the general quality of the raw hybridisation data, observing a moderately worse performance than what had been previously observed for array-based resequencing of the less complex *Arabidopsis thaliana* genome. Moreover, we observed an inverse correlation between the hybridisation performance and the evolutionary distance to the reference Nipponbare. The use of LR-PCR for amplification of the target DNA with primers complementary to the reference sequence is one potential factor that almost certainly affected data quality. With such an experimental design, amplification can fail if complementarity of primers to the target genome is compromised due to polymorphisms in the target sequence. For the *Arabidopsis* project, a different experimental design – whole genome DNA amplification – had been chosen that allowed unbiased sampling.

The hybridisation data was preprocessed by quantile normalisation to remove differences due to experimental variation. As the normalisation was performed on a per *amplicon pool* basis, the normalisation was limited to incorporate intensity variations caused by differences in experimental conditions between the pools of samples hybridised to arrays. As an extension, one could apply the quantile normalisation method at the amplicon level. The idea is motivated by the observation that sequence context and amplicon length influence amplification efficiency. This leads to varying amplicon concentrations and thus to different hybridisation intensity levels. A second idea to further improve comparability of the data is to take into account that different sequence properties of the oligonucleotide probes on the arrays also cause intensity variability. This probe sequence effect can be partially removed by using an approach similar to the one described in Zeller et al. (2007b).

We have also shown that repetitive 25-mers have a significant influence on hybridisation intensity, a potential source of error for downstream analyses. To annotate potential cross-hybridisation sites, we transversed the rice reference genome in search of 25-mers that occur multiple times. We restricted the match pairs to those between a 25-mer probe on an array and a 25-mer contained in one of the amplicons in the corresponding amplicon pool. As the analysis of repetitive 25-mers is based on the reference Nipponbare, sequences that occur in the genome of another strain could not be taken into account. Thus, cross-hybridisation sites specific to one strain likely remained undetected. To address this problem, a further improvement of the k-mer analysis could include information on common repeats from the rice repeat database published by Ouyang and Buell (2004). This database comprises simple repeats, mobile elements such as transposons and retrotransposons, and repeats in the centromers and telomeric sequences detected by biological experiments and bioinformatic analyses. Mobile elements make up a high proportion of the repetitive DNA in rice, and the content of these

elements can vary widely even between individuals of the same subspecies.

For training and evaluation of our SNP calling method, a collection of polymorphisms identified by dideoxy sequencing was annotated. This set contained 9,499 SNPs within 1,755,395 sequenced positions, resulting in an estimate of one SNP per 5,000 bases. We undertook various analyses to assess the quality of the dideoxy sequencing set, observing a higher label confidence with higher sequence quality scores.

Finally, having established all the prerequisites for the application of the machine learning method, we trained Support Vector Machines to learn to discriminate between true SNP and non-SNP positions based on the array hybridisation data. The trained machines were then used to generate predictions for all positions interrogated by the tiling arrays. Our method outperformed the MB method developed by Perlegen Sciences, achieving 20.9% recall at a FDR of 8.3%. We predicted 1,366,243 SNPs across all strains in total, of which 55.9% overlapped with the MB SNP calls. However, the performance of both SNP calling methods was considerably poorer than the performance previously achieved for SNP detection in the *Arabidopsis thaliana* strains. As many factors differed between the rice and *Arabidopsis* resequencing project, one can only guess at potential reasons for the weaker performance.

Apart from SNP discovery, array hybridisation data can also be exploited to detect deletions and regions of high dissimilarity. In the future we plan to apply a previously developed method to predict such highly polymorphic regions (see Zeller et al., 2007a). A SNP at a certain position causes a mismatch in probes at the positions nearby and therefore decreases the hybridisation intensities for these probes. In highly polymorphic regions, closely spaced SNPs result in a pattern of extended and poor hybridisation. This leads to a drastic underestimation of the SNP density. To uncover tracts of such reduced hybridisation intensity, a state-model based approach as proposed in Zeller et al. (2007a) can be applied to predict the segmentation into conserved and polymorphic regions. The application of this method to the *Arabidopsis* resequencing data showed that the fraction of called polymorphism can dramatically be increased in this way.

In summary, the analysis of the resequencing data from diverse rice strains will provide insights into the polymorphic patterns in a very important crop species. In this Diplom thesis, the application of computational approaches for SNP discovery resulted in the identification of SNPs in non-repetitive parts of the genome. At the time when the rice resequencing project was initiated, high-density microarrays were the method of choice for resequencing since they were the most cost-effective technology to resequence a larger number of individuals in parallel. Notwithstanding the benefits of this technology, microarrays have various limitations. For whole-genome studies, a reference genome sequence has to be available to serve as a template for the array probes. Secondly, many different factors such as cross-hybridisation can introduce distortions of the hybridisation signal and affect the reliability of the intensity measurements. Altogether, sequence variation is measured via hybridisation in an indirect way, requiring various preprocessing steps as I illustrated in this thesis. Due to its limitations, array-based resequencing is likely to be replaced or at least complemented in the near future by sequencing-by-synthesis methods, which are also very cost-effective (Hutchison, 2007). However, many of the conceptual ideas developed in this thesis and in related work will be applicable to these new technologies, which currently generate very short sequence tags not much different in lengths from the oligonucleotides employed on resequencing arrays.

References

Affymetrix. Microarray Manufacturing. Website, September 2007. URL `http://www.affymetrix.com/technology/manufacturing/index.affx`.

Ben M. Bolstad, Rafael A. Irizarry, Magnus Åstrand, and Terry P. Speed. A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Variance and Bias. *Bioinformatics*, 19:185–193, 2003.

James K. Bonfield and Rodger Staden. ZTR: A New Format for DNA Sequence Trace Data. *Bioinformatics*, 18:3–10, 2002.

Conrad J. Burden, Yvonne Pittelkow, and Susan R. Wilson. Adsorption Models of Hybridization and Post-hybridization Behaviour on Oligonucleotide Microarrays. *Journal of Physics*, 18(23):5545–5565, 2006.

Christopher J. C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, 2:121–167, 1998.

Ana Caicedo, Scott Williamson, Ryan D. Hernandez, Adam Boyko, Adi Fledel-Alon, Thomas York, Nicholas Polato, Ken Olsen, Rasmus Nielsen, Susan McCouch, Carlos D. Bustamante, and Michael D. Purugganan. Genome-wide Patterns of Nucleotide Polymorphism in Domesticated Rice. *PLoS Genetics*, preprint:e163.eor, August 2007.

Richard M. Clark, Gabriele Schweikert, Christopher Toomajian, Stephan Ossowski, Georg Zeller, Paul Shinn, Norman Warthmann, Tina T. Hu, Glenn Fu, David A. Hinds, Huaming Chen, Kelly A. Frazer, Daniel H. Huson, Bernhard Schölkopf, Magnus Nordborg, Gunnar Rätsch, Joseph R. Ecker, and Detlef Weigel. Common Sequence Polymorphisms Shaping Genetic Diversity in *Arabidopsis thaliana*. *Science*, 317:338–42, July 2007.

Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms*. MIT Press, Cambridge, MA, 2001.

David J. Cutler, Michael E. Zwick, Minerva M. Carrasquillo, Christopher T. Yohn, Katherine P. Tobin, Carl Kashuk, Debra J. Mathews, Nila A. Shah, Evan E. Eichler, Janet A. Warrington, and Aravinda Chakravarti. High-throughput Variation Detection and Genotyping Using Microarrays. *Genome Research*, 11:1913–25, November 2001.

Brent Ewing and Phil Green. Base-calling of Automated Sequencer Traces Using Phred. II. Error Probabilities. *Genome Research*, 8:186–94, March 1998.

Kelly A. Frazer, Eleazar Eskin, Hyun Min Kang, Molly A. Bogue, David A. Hinds, Erica J. Beilharz, Robert V. Gupta, Julie Montgomery, Matt M. Morenzoni, Geoffrey B. Nilsen, Charit L. Pethiyagoda, Laura L. Stuve, Frank M. Johnson, Mark J. Daly, Claire M. Wade, and David R. Cox. A Sequence-based Variation Map of 8.27 Million SNPs in Inbred Mouse Strains. *Nature*, 448:1050–3, August 2007.

Amanda J. Garris, Thomas H. Tai, Jason Coburn, Steve Kresovich, and Susan McCouch. Genetic Structure and Diversity in *Oryza sativa* L. *Genetics*, 169:1631–8, March 2005.

Stephen A. Goff, Darrell Ricke, Tien-Hung Lan, Gernot Presting, Ronglin Wang, Molly Dunn, Jane Glazebrook, Allen Sessions, Paul Oeller, Hemant Varma, David Hadley, Don Hutchison, Chris Martin, Fumiaki Katagiri, B. Markus Lange, Todd Moughamer, Yu Xia, Paul Budworth, Jingping

Zhong, Trini Miguel, Uta Paszkowski, Shiping Zhang, Michelle Colbert, Wei lin Sun, Lili Chen, Bret Cooper, Sylvia Park, Todd Charles Wood, Long Mao, Peter Quail, Rod Wing, Ralph Dean, Yeisoo Yu, Andrey Zharkikh, Richard Shen, Sudhir Sahasrabudhe, Alun Thomas, Rob Cannings, Alexander Gutin, Dmitry Pruss, Julia Reid, Sean Tavtigian, Jeff Mitchell, Glenn Eldredge, Terri Scholl, Rose Mary Miller, Satish Bhatnagar, Nils Adey, Todd Rubano, Nadeem Tusneem, Rosann Robinson, Jane Feldhaus, Teresita Macalma, Arnold Oliphant, and Steven Briggs. A Draft Sequence of the Rice Genome (*Oryza sativa* L. ssp. *japonica*). *Science*, 296:92–100, April 2002.

Joseph G. Hacia. Resequencing and Mutational Analysis Using Oligonucleotide Microarrays. *Nature Genetics*, 21:42–7, 1999.

Bin Han and Yongbiao Xue. Genome-wide Intraspecific DNA-sequence Variations in Rice. *Current Opinion in Plant Biology*, 6:134–8, April 2003.

David A. Hinds, Laura L. Stuve, Geoffrey B. Nilsen, Eran Halperin, Eleazar Eskin, Dennis G. Ballinger, Kelly A. Frazer, and David R. Cox. Whole-genome Patterns of Common DNA Variation in Three Human Populations. *Science*, 307:1072–9, February 2005.

Clyde A. Hutchison. DNA Sequencing: Bench to Bedside and Beyond. *Nucleic Acids Research*, doi: 10.1093/nar/gkm688, September 2007.

International Rice Genome Sequencing Project. The Map-based Sequence of the Rice Genome. *Nature*, 436:793–800, August 2005.

International Rice Research Institute. Rice Statistics. Website, September 2007a. URL `http://www.irri.org/science/ricestat/index.asp`.

International Rice Research Institute. IRRI Photo Bank. Website, September 2007b. URL `http://www.ricephotos.org`.

Mazen W. Karaman, Susan Groshen, Chi-Chiang Lee, Brian L. Pike, and Joseph G. Hacia. Comparisons of Substitution, Insertion and Deletion Probes for Resequencing and Mutational Analysis Using Oligonucleotide Microarrays. *Nucleic Acids Research*, 33:e33, 2005.

W. James Kent. BLAT – the BLAST-like Alignment Tool. *Genome Research*, 12:656–64, April 2002.

Stefan Kurtz, Adam Phillippy, Arthur L. Delcher, Michael Smoot, Martin Shumway, Corina Antonescu, and Steven L. Salzberg. Versatile and Open Software for Comparing Large Genomes. *Genome Biology*, 5:R12, 2004.

Inhan Lee, Alan A. Dombkowski, and Brian D. Athey. Guidelines for Incorporating Non-perfectly Matched Oligonucleotides into Target-specific Hybridization Probes for a DNA Microarray. *Nucleic Acids Research*, 32:681–90, 2004.

Xiaohui Liu, Tingting Lu, Shuliang Yu, Ying Li, Yuchen Huang, Tao Huang, Lei Zhang, Jingjie Zhu, Qiang Zhao, Danlin Fan, Jie Mu, Yingying Shangguan, Qi Feng, Jianping Guan, Kai Ying, Yu Zhang, Zhixin Lin, Zongxiu Sun, Qian Qian, Yuping Lu, and Bin Han. A Collection of 10,096 *Indica* Rice Full-length cDNAs Reveals Highly Expressed Sequence Divergence Between *Oryza sativa indica* and *japonica* Subspecies. *Plant Molecular Biology*, May 2007.

Harvey Lodish, Arnold Berk, Paul Matsudaira, Chris A. Kaiser, Monty Krieger, Matthew P. Scott, S. Lawrence Zipursky, and James Darnell. *Molecular Cell Biology*. W. H. Freeman and Company, New York, 2004.

Kenneth L. McNally, Richard Bruskiewich, David Mackill, C. Robin Buell, Jan E. Leach, and Hei Leung. Sequencing Multiple and Diverse Rice Varieties. Connecting Whole-genome Variation with Phenotypes. *Plant Physiology*, 141:26–31, May 2006.

Todd C. Mockler, Simon Chan, Ambika Sundaresan, Huaming Chen, Steven E. Jacobsen, and Joseph R. Ecker. Applications of DNA Tiling Arrays for Whole-genome Analysis. *Genomics*, 85:1–15, 2005.

Lisa Monna, Rieko Ohta, Haruka Masuda, Akiko Koike, and Yuzo Minobe. Genome-wide Searching of Single Nucleotide Polymorphisms Among Eight Distantly and Closely Related Rice Cultivars (*Oryza sativa* L.) and a Wild Accession (*Oryza rufipogon* Griff.). *DNA Research*, 13:43–51, April 2006.

Magnus Nordborg, Tina T. Hu, Yoko Ishino, Jinal Jhaveri, Christopher Toomajian, Honggang Zheng, Erica Bakker, Peter Calabrese, Jean Gladstone, Rana Goyal, Mattias Jakobsson, Sung Kim, Yuri Morozov, Badri Padhukasahasram, Vincent Plagnol, Noah A. Rosenberg, Chitiksha Shah, Jeffrey D. Wall, Jue Wang, Keyan Zhao, Theodore Kalbfleisch, Vincent Schulz, Martin Kreitman, and Joy Bergelson. The Pattern of Polymorphism in *Arabidopsis thaliana*. *PLoS Biology*, 3:e196, July 2005.

Shu Ouyang and C. Robin Buell. The TIGR Plant Repeat Databases: A Collective Resource for the Identification of Repetitive Sequences in Plants. *Nucleic Acids Research*, 32:D360–3, 2004.

Perlegen Sciences. Rice SNP Discovery Final Technical Report. Technical report, Perlegen Sciences, June 2007.

Graham Ramsay. DNA Chips: State-of-the Art. *Nature Biotechnology*, 16:40–4, 1998.

Willem A. Rensink and C. Robin Buell. *Arabidopsis* to Rice. Applying Knowledge from a Weed to Enhance Our Understanding of a Crop Species. *Plant Physiology*, 135:622–9, June 2004.

Eric E. Schadt, Cheng Li, Beverley Ellis, and Wing H. Wong. Feature Extraction and Normalization Algorithms for High-density Oligonucleotide Gene Expression Array Data. *Journal of Cellular Biochemistry*, Suppl 37:120–5, 2001.

Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

Ying-Jia Shen, Hua Jiang, Jian-Peng Jin, Zai-Bao Zhang, Biao Xi, You-Yu He, Guan Wang, Chen Wang, Lily Qian, Xiang Li, Qing-Bo Yu, Hui-Juan Liu, De-Hui Chen, Jian-Hua Gao, Hai Huang, Tie-Liu Shi, and Zhong-Nan Yang. Development of Genome-Wide DNA Polymorphism Database for Map-Based Cloning of Rice Genes. *Plant Physiology*, 135:1198–1205, July 2004.

Sören Sonnenburg, Gunnar Rätsch, Christin Schäfer, and Bernhard Schölkopf. Large Scale Multiple Kernel Learning. *J. Mach. Learn. Res*, 7:1531–1565, 2006.

Sören Sonnenburg, Gabriele Schweikert, Petra Philips, Jonas Behr, and Gunnar Rätsch. Accurate Splice Site Prediction. Submitted, 2007.

Megan Sweeney and Susan McCouch. The Complex History of the Domestication of Rice. *Annals of Botany*, 100:951–57, July 2007.

Ann-Christine Syvänen. Accessing Genetic Variation: Genotyping Single Nucleotide Polymorphisms. *Nature Reviews Genetics*, 2:930–42, December 2001.

The Institute for Genome Research. Rice Genome Facts. Website, September 2007. URL `http://www.tigr.org/tdb/e2k1/osa1/riceInfo/info.shtml#Genes`.

The Rice Annotation Project. Curated Genome Annotation of *Oryza sativa* ssp. *japonica* and Comparative Genome Analysis with *Arabidopsis thaliana*. *Genome Research*, 17:175–83, February 2007.

Julie D. Thompson, Desmond G. Higgins, and Toby J. Gibson. CLUSTAL W: Improving the Sensitivity of Progressive Multiple Sequence Alignment Through Sequence Weighting, Position-specific Gap Penalties and Weight Matrix Choice. *Nucleic Acids Research*, 22:4673–80, November 1994.

Vladimir N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, Inc., 1998.

Norman Warthmann, Joffrey Fitz, and Detlef Weigel. MSQT for Choosing Snp Assays from Multiple DNA Alignments. *Bioinformatics*, September 2007.

Detlef Weigel and Magnus Nordborg. Natural Variation in Arabidopsis. How Do We Find the Causal Genes? *Plant Physiology*, 138:567–8, June 2005.

Wikipedia. Rice. Website, September 2007. URL `http://en.wikipedia.org/wiki/Rice`.

Jun Yu, Songnian Hu, Jun Wang, Gane Ka-Shu Wong, Songgang Li, Bin Liu, Yajun Deng, Li Dai, Yan Zhou, Xiuqing Zhang, Mengliang Cao, Jing Liu, Jiandong Sun, Jiabin Tang, Yanjiong Chen, Xiaobing Huang, Wei Lin, Chen Ye, Wei Tong, Lijuan Cong, Jianing Geng, Yujun Han, Lin Li, Wei Li, Guangqiang Hu, Xiangang Huang, Wenjie Li, Jian Li, Zhanwei Liu, Long Li, Jianping Liu, Qiuhui Qi, Jinsong Liu, Li Li, Tao Li, Xuegang Wang, Hong Lu, Tingting Wu, Miao Zhu, Peixiang Ni, Hua Han, Wei Dong, Xiaoyu Ren, Xiaoli Feng, Peng Cui, Xianran Li, Hao Wang, Xin Xu, Wenxue Zhai, Zhao Xu, Jinsong Zhang, Sijie He, Jianguo Zhang, Jichen Xu, Kunlin Zhang, Xianwu Zheng, Jianhai Dong, Wanyong Zeng, Lin Tao, Jia Ye, Jun Tan, Xide Ren, Xuewei Chen, Jun He, Daofeng Liu, Wei Tian, Chaoguang Tian, Hongai Xia, Qiyu Bao, Gang Li, Hui Gao, Ting Cao, Juan Wang, Wenming Zhao, Ping Li, Wei Chen, Xudong Wang, Yong Zhang, Jianfei Hu, Jing Wang, Song Liu, Jian Yang, Guangyu Zhang, Yuqing Xiong, Zhijie Li, Long Mao, Chengshu Zhou, Zhen Zhu, Runsheng Chen, Bailin Hao, Weimou Zheng, Shouyi Chen, Wei Guo, Guojie Li, Siqi Liu, Ming Tao, Jian Wang, Lihuang Zhu, Longping Yuan, and Huanming Yang. A Draft Sequence of the Rice Genome (*Oryza sativa* L. ssp. *indica*). *Science*, 296:79–92, April 2002.

Georg Zeller, Richard M. Clark, Korbinian Schneeberger, Anja Bohlen, Detlef Weigel, and Gunnar Rätsch. Detecting Polymorphic Regions in the *Arabidopsis thaliana* Genome with Resequencing Microarrays. Submitted, 2007a.

Georg Zeller, Stefan R. Henz, Sascha Laubinger, Detlef Weigel, and Gunnar Rätsch. Transcript Normalisation and Segmentation of Tiling Array Data. In *Pacific Symposium on Biocomputing Proceedings*, volume 12, pages 527–38, September 2007b.

Yiping Zhan and David Kulp. Model-P: A Basecalling Method for Resequencing Microarrays of Diploid Samples. *Bioinformatics*, 21 Suppl 2:ii182–9, September 2005.

Appendix

A.1 Figures

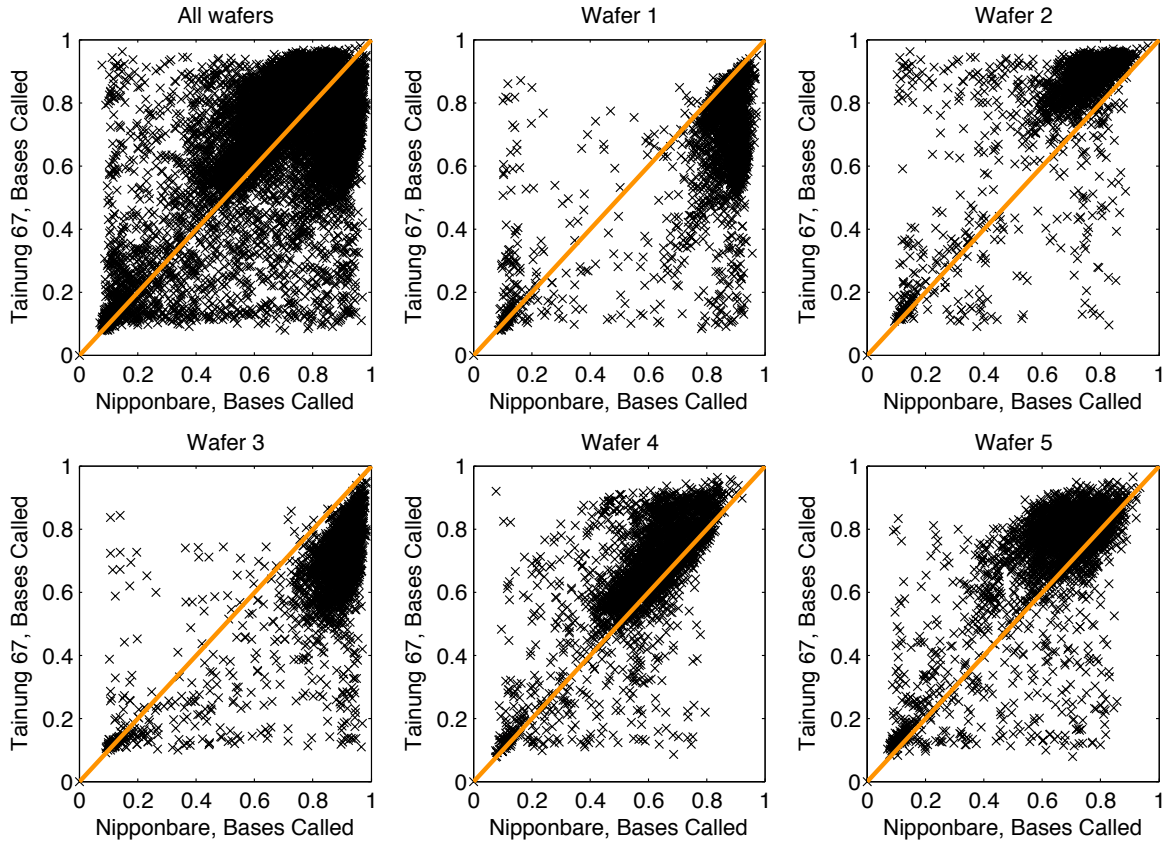


Figure A.1: Scatter plots of the fractions of bases called as reference for Nipponbare versus those for the *japonica* strain Tainung 67. The analysis was made for all wafers together and on a per wafer basis to reveal wafer-dependent effects.

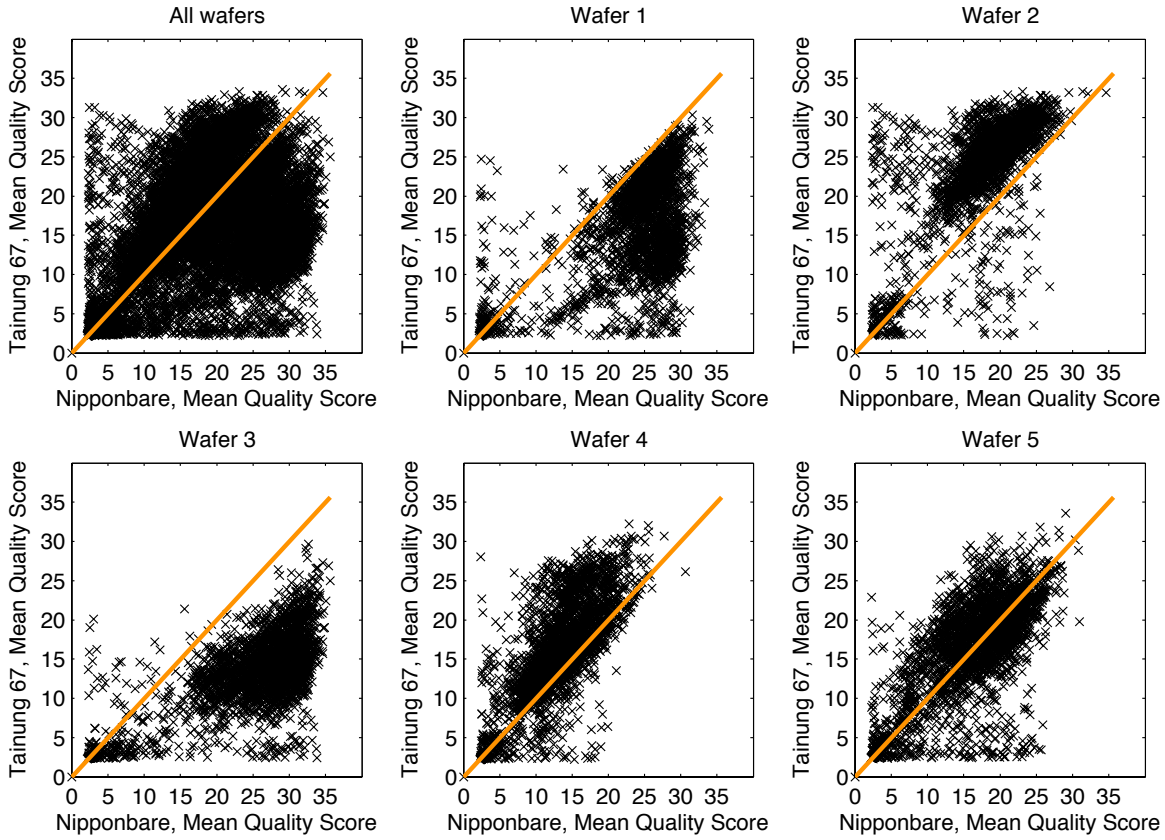


Figure A.2: Scatter plots of mean quality scores for Nipponbare versus the *japonica* strain Tainung 67 are visualised for each wafer separately.

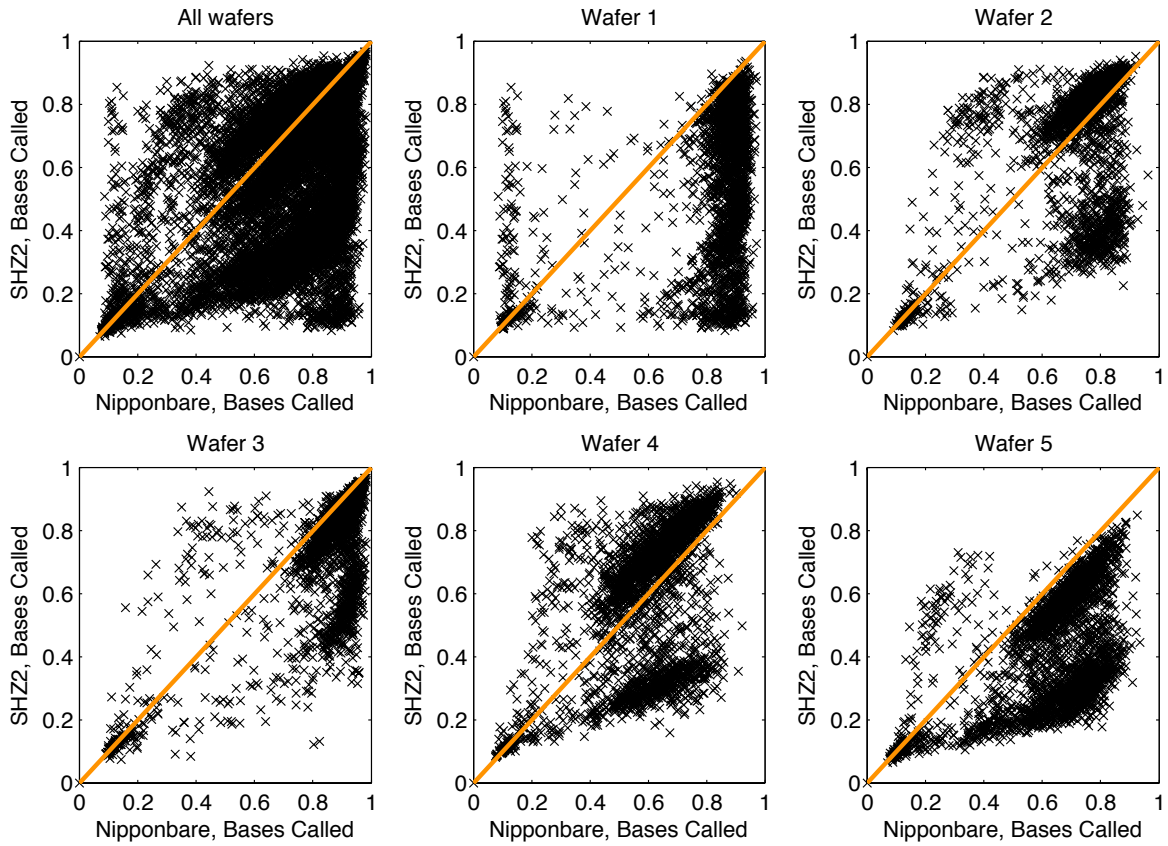


Figure A.3: Scatter plots of the fractions of bases called as reference for Nipponbare versus those for the *indica* strain SHZ2. The analysis was made for all wafers together and on a per wafer basis to reveal wafer-dependent effects.

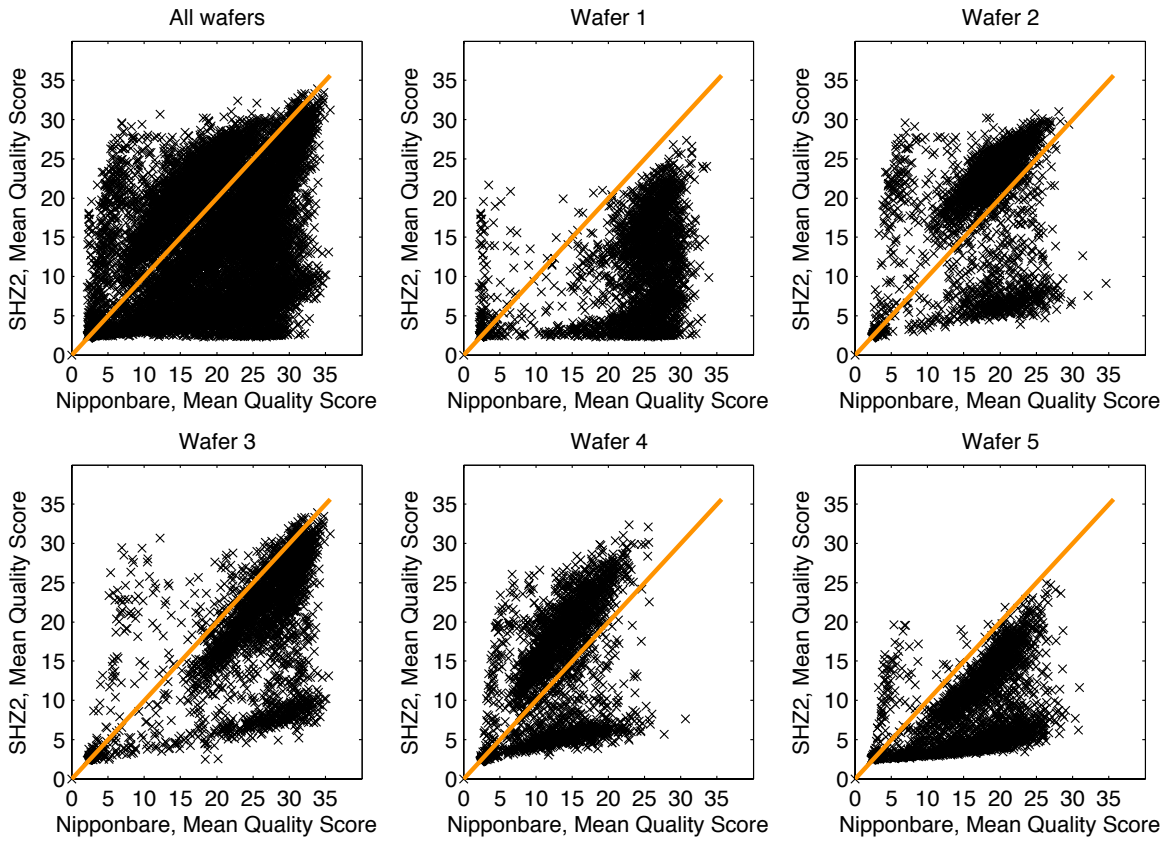
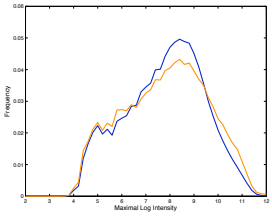
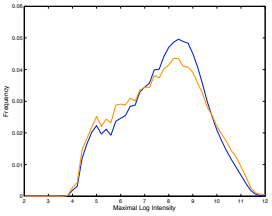


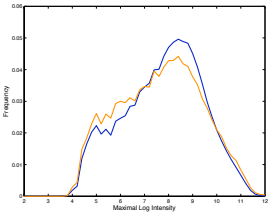
Figure A.4: Scatter plots of mean quality scores for Nipponbare versus the *indica* strain SHZ2 are visualised for each wafer separately.



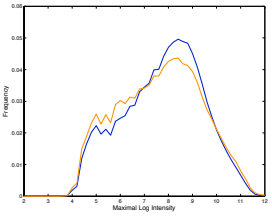
(a) Mismatch at Pos. 3



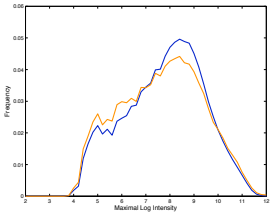
(b) Mismatch at Pos. 5



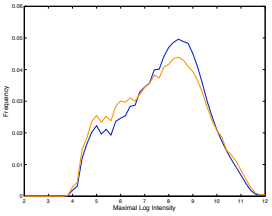
(c) Mismatch at Pos. 9



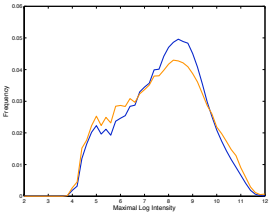
(d) Mismatch at Pos. 11



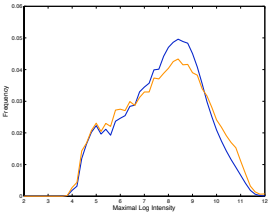
(e) Mismatch at Pos. 14



(f) Mismatch at Pos. 17

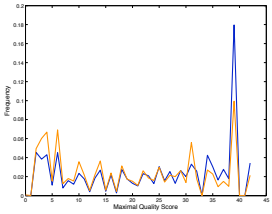


(g) Mismatch at Pos. 21

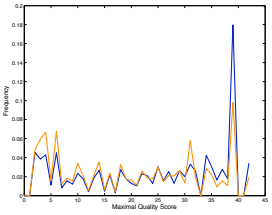


(h) Mismatch at Pos. 23

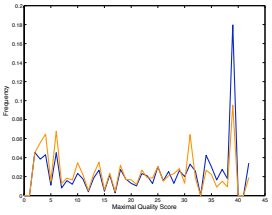
Figure A.5: Effects of inexact 25-mer matches on hybridisation at different mismatch positions. The distribution of the maximal hybridisation intensities are illustrated for all unique probes (blue) and for probes with inexact 25-mer matches (orange).



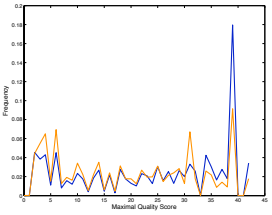
(a) Mismatch at Pos. 3



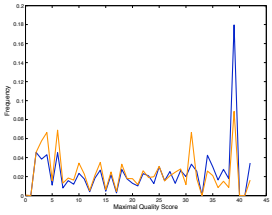
(b) Mismatch at Pos. 5



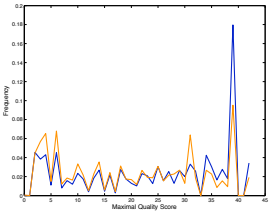
(c) Mismatch at Pos. 9



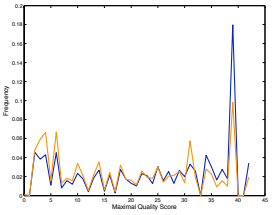
(d) Mismatch at Pos. 11



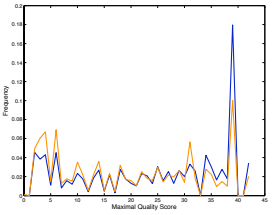
(e) Mismatch at Pos. 14



(f) Mismatch at Pos. 17

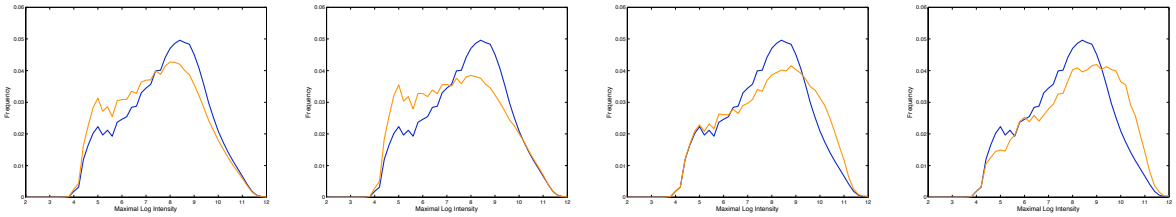


(g) Mismatch at Pos. 21

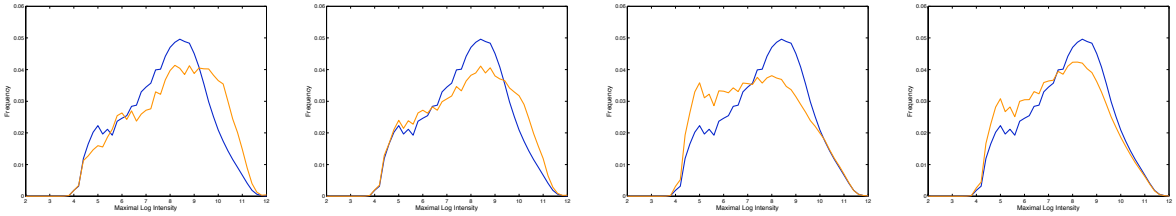


(h) Mismatch at Pos. 23

Figure A.6: Effects of inexact 25-mer matches on hybridisation at different mismatch positions. The histograms of maximal quality scores are shown for positions with (orange) and without (blue) inexact 25-mer matches.

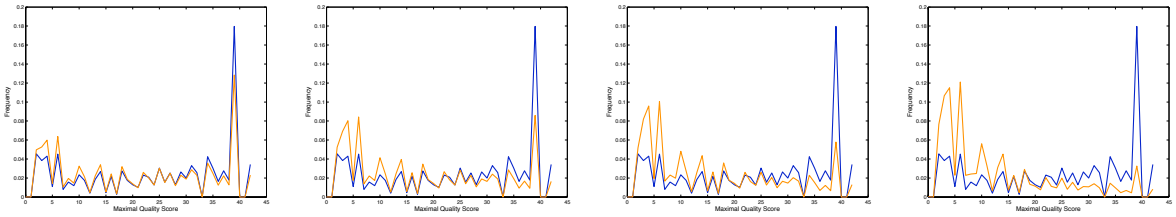


(a) Bulge at Position 3 (b) Bulge at Position 5 (c) Bulge at Position 8 (d) Bulge at Position 12

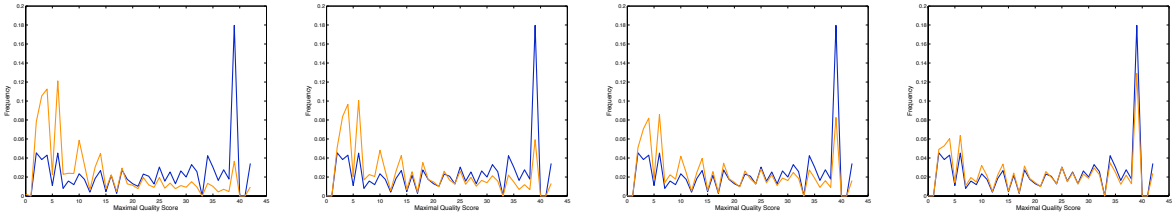


(e) Bulge at Position 15 (f) Bulge at Position 19 (g) Bulge at Position 22 (h) Bulge at Position 24

Figure A.7: Effects of bulged 25-mer matches on hybridisation for bulges at different positions. The distribution of the maximal hybridisation intensities are illustrated for all unique probes (blue) and in orange for probes with bulged 25-mer matches (orange).



(a) Bulge at Position 3 (b) Bulge at Position 5 (c) Bulge at Position 8 (d) Bulge at Position 12



(e) Bulge at Position 15 (f) Bulge at Position 19 (g) Bulge at Position 22 (h) Bulge at Position 24

Figure A.8: Effects of bulged 25-mer matches on hybridisation for bulges at different positions. The histograms of maximal quality scores are shown for positions with (orange) and without (blue) bulged 25-mer matches.

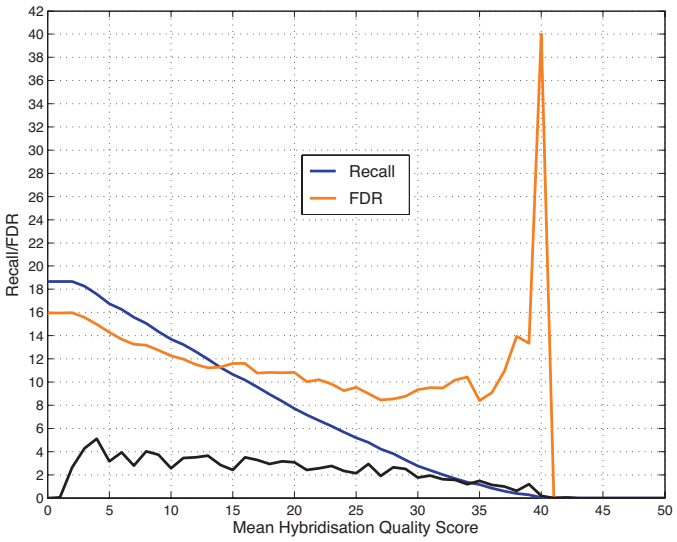


Figure A.9: ML performance dependency on hybridisation quality scores. The figure shows the distribution of hybridisation quality scores (black), the recall (blue) and the FDR (orange) as a function of the mean hybridisation quality score cut-off.

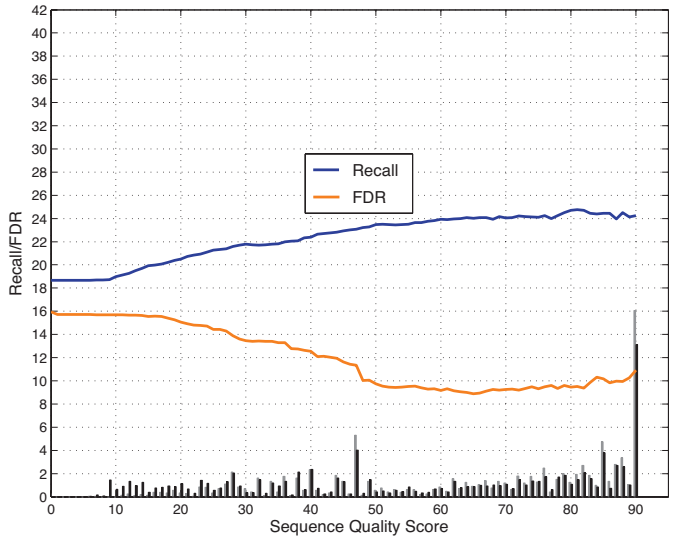


Figure A.10: ML performance dependency on sequencing quality scores. The distribution of the sequence quality score at all predicted postions and at all labelled SNP positions are given in dark and light grey, respectively. Recall (blue) and FDR (orange) are shown as a function of the sequence quality score cut-off.

A.2 Tables

Table A.1: For each strain the percentage of positions for which the base call for the forward strand was identical with the base call on the reverse strand are illustrated on a per wafer basis.

Strain	Wafer						
	0	1	2	3	4	5	All
Aswina	94.6	73.6	69.7	80.9	71.1	75.9	74.3
Azucena	93.6	74.3	76.3	73.5	69.9	72.5	73.4
Cypress	88.4	81.4	78.2	72.8	77.2	73.8	76.7
Dom Sufid	90.3	74.8	81.7	77.1	82.1	69.9	77.2
Dular	91.1	71.0	74.6	74.1	75.7	69.8	73.1
FR 13A	88.7	77.8	73.2	69.6	68.4	70.6	72.0
IR 64	92.0	75.3	72.3	75.7	63.5	71.1	71.7
LTH	96.2	74.7	82.2	73.7	77.4	73.8	76.4
M 202	92.1	75.6	79.6	81.0	71.2	75.2	76.6
Minghui 63	86.2	76.6	69.9	72.2	63.7	67.6	70.1
Moroberekan	93.9	81.6	75.5	77.9	71.6	71.3	75.6
N 22	91.9	79.8	74.8	81.9	69.4	74.5	76.1
Nipponbare	91.0	84.2	76.2	86.5	66.2	68.1	76.3
Pokkali	91.0	73.3	75.5	75.3	61.8	69.2	71.1
Rayada	87.0	72.2	78.4	70.6	76.1	64.1	72.3
Sadu Cho	88.7	75.5	77.4	71.1	69.6	67.6	72.3
SHZ2	90.1	66.7	74.9	78.8	66.8	52.7	68.1
Swarna	91.8	67.5	77.0	67.9	73.2	65.4	70.3
Tainung 67	92.5	74.3	85.5	72.9	73.4	75.7	76.4
Zhenshan 97	91.1	75.0	81.3	71.0	81.0	71.5	76.0
Average	91.1	75.3	76.7	75.2	71.5	70.0	73.8

Table A.2: For each strain the percentage of positions for which the base calls on both strands were identical and correspond to the reference base are illustrated on a per wafer basis.

Strain	Wafer						
	0	1	2	3	4	5	All
Aswina	93.1	68.3	61.3	75.6	59.9	69.3	67.0
Azucena	92.3	68.4	71.0	65.1	58.9	63.1	65.4
Cypress	84.5	77.7	72.9	64.8	70.1	66.3	70.5
Dom Sufid	87.7	69.7	77.6	70.9	76.4	60.8	71.2
Dular	87.4	64.3	67.9	66.1	67.0	60.4	65.2
FR 13A	84.6	71.9	67.0	61.3	57.5	61.0	63.8
IR 64	89.9	69.1	64.0	67.8	48.8	61.9	62.5
LTH	95.4	69.6	78.9	67.1	72.6	65.7	70.9
M 202	89.1	69.9	75.2	77.0	61.6	69.3	70.7
Minghui 63	80.8	70.4	61.8	62.5	52.6	56.7	60.9
Moroberekan	91.9	78.5	70.0	71.7	65.3	63.5	69.9
N 22	89.0	75.8	69.5	77.6	62.5	67.8	70.7
Nipponbare	88.7	81.5	70.9	83.8	58.7	62.4	71.5
Pokkali	88.7	65.6	68.6	67.6	47.8	60.6	62.2
Rayada	83.9	67.0	73.2	61.9	69.7	52.3	64.9
Sadu Cho	86.0	69.3	71.5	61.6	61.3	55.8	64.0
SHZ2	86.9	55.2	68.8	74.0	57.7	41.7	59.6
Swarna	89.1	58.6	71.9	59.4	64.5	55.8	62.1
Tainung 67	90.0	68.3	82.1	65.4	65.4	69.1	70.1
Zhenshan 97	88.2	69.0	77.8	63.2	75.6	62.4	69.7
Average	88.4	69.4	71.1	68.2	62.7	61.3	66.2

Table A.3: Match pairs for the four k-mer match types.

Pool	Exact	Inexact	Short	Bulged	Total
0	9,184	30,863	16,563	25,472	82,082
1	147,037	319,535	136,040	301,614	904,226
2	169,989	345,234	145,166	355,523	1,015,912
3	130,933	259,973	114,956	269,808	775,670
4	130,246	271,107	125,254	284,325	810,932
5	116,903	251,178	109,103	242,023	719,207
6	297,225	552,184	211,788	566,562	1,627,759
7	114,079	233,599	106,270	219,959	673,907
8	164,417	378,146	157,429	337,746	1,037,738
9	208,815	484,757	196,781	417,443	1,307,796
10	226,909	462,526	186,347	448,346	1,324,128
11	202,267	410,442	170,101	419,989	1,202,799
12	195,452	417,338	175,978	401,695	1,190,463
13	199,730	376,793	155,617	391,532	1,123,672
14	286,339	560,199	213,845	534,451	1,594,834
15	403,154	674,116	257,198	767,439	2,101,907
Total	3,002,679	6,027,990	2,478,436	5,983,927	17,493,032

Table A.4: Repetitive positions for the four k-mer match types. The percentage is given in squared brackets.

Pool	Exact	Inexact	Short	Bulged
0	7,839 [2.07]	11,030 [2.91]	5,556 [1.47]	5,515 [1.45]
1	132,846 [2.04]	137,059 [2.10]	52,067 [0.80]	82,767 [1.27]
2	143,470 [2.20]	142,902 [2.19]	54,588 [0.84]	88,932 [1.37]
3	110,337 [1.59]	107,191 [1.55]	42,604 [0.61]	68,599 [0.99]
4	112,932 [1.73]	108,769 [1.67]	45,420 [0.70]	71,846 [1.10]
5	105,538 [1.62]	110,122 [1.69]	41,798 [0.64]	64,885 [1.00]
6	248,903 [3.60]	231,646 [3.35]	81,855 [1.18]	151,964 [2.20]
7	89,400 [1.37]	96,255 [1.48]	39,307 [0.60]	56,349 [0.86]
8	133,214 [2.05]	157,593 [2.42]	58,930 [0.91]	85,565 [1.31]
9	168,122 [2.43]	198,547 [2.87]	74,367 [1.08]	107,555 [1.56]
10	178,206 [2.74]	193,503 [2.97]	70,752 [1.09]	113,228 [1.74]
11	173,088 [2.66]	176,804 [2.72]	65,229 [1.00]	109,883 [1.69]
12	170,121 [2.46]	177,587 [2.57]	66,987 [0.97]	106,107 [1.53]
13	179,950 [2.77]	164,650 [2.53]	60,078 [0.92]	108,596 [1.67]
14	239,219 [3.68]	238,392 [3.67]	82,533 [1.27]	144,510 [2.22]
15	334,995 [4.85]	278,941 [4.03]	97,984 [1.42]	203,103 [2.94]
Total	2,528,180 [2.53]	2,530,991 [2.53]	940,055 [0.94]	1,569,404 [1.57]

Table A.5: Shown is the number of MB SNP calls broken down according to different filter criteria. We grouped the strains with respect to their population affiliation. The two performance measures [recall, FDR] are given in % in squared brackets. N/A denotes FDRs which could not be assessed as no predictions were made.

Strain	No Filter	Repetitive 25-mer Matches	Exact 25-mer Matches	Excluding Positions with Dominating 25-mer Matches	Hybridisation Quality Score ≤ 8	Sequencing Quality Score ≤ 48	Repetitive Fragments (90% Identity)	Singleton SNPs
LTH	5 [11.9, 00.0]	5 [11.9, 00.0]	5 [11.9, 00.0]	5 [11.9, 00.0]	5 [11.9, 00.0]	4 [66.7, 00.0]	4 [16.7, 00.0]	0 [00.0, N/A]
M 202	6 [09.1, 16.7]	6 [09.1, 16.7]	6 [09.1, 16.7]	6 [09.1, 16.7]	6 [09.1, 16.7]	4 [33.3, 00.0]	6 [12.2, 16.7]	3 [07.7, 00.0]
Tainung 67	4 [22.2, 50.0]	3 [22.2, 33.3]	4 [22.2, 50.0]	3 [22.2, 33.3]	3 [22.2, 33.3]	3 [66.7, 33.3]	4 [28.6, 50.0]	3 [33.3, 33.3]
Azucena	22 [07.7, 22.7]	22 [07.7, 22.7]	22 [07.7, 22.7]	22 [07.7, 22.7]	19 [06.8, 21.1]	15 [10.7, 20.0]	19 [08.8, 21.1]	20 [09.5, 20.0]
Cypress	37 [08.0, 10.8]	35 [07.8, 08.6]	36 [08.0, 08.3]	36 [07.8, 11.1]	36 [08.0, 08.3]	24 [11.1, 12.5]	34 [08.1, 08.8]	31 [10.0, 12.9]
Moroberekan	19 [05.9, 21.1]	17 [05.5, 17.6]	18 [05.9, 16.7]	18 [05.5, 22.2]	18 [05.9, 16.7]	11 [06.3, 09.1]	19 [06.5, 21.1]	12 [06.8, 08.3]
Dom Sufid	13 [06.1, 61.5]	11 [06.1, 54.5]	12 [06.1, 58.3]	12 [06.1, 58.3]	9 [06.1, 44.4]	6 [04.7, 66.7]	12 [06.4, 58.3]	12 [06.8, 66.7]
Dular	103 [13.0, 17.5]	94 [12.1, 16.0]	97 [12.4, 16.5]	98 [12.5, 16.3]	99 [12.7, 16.2]	63 [13.8, 12.7]	88 [13.8, 17.0]	58 [11.3, 15.5]
FR 13A	108 [13.7, 13.0]	103 [13.2, 11.7]	105 [13.4, 12.4]	105 [13.4, 12.4]	105 [13.2, 13.3]	68 [15.7, 08.8]	100 [14.3, 11.0]	78 [13.2, 11.5]
N 22	125 [11.8, 19.2]	117 [11.4, 17.1]	120 [11.5, 18.3]	120 [11.6, 17.5]	121 [11.7, 17.4]	81 [14.3, 09.9]	112 [12.3, 15.2]	89 [12.7, 19.1]
Rayada	146 [15.0, 07.5]	141 [14.5, 07.1]	142 [14.6, 07.0]	142 [14.6, 07.0]	144 [15.0, 06.2]	99 [18.0, 04.0]	133 [16.3, 06.0]	101 [16.4, 06.9]
Aswina	71 [14.0, 18.3]	68 [14.0, 14.7]	69 [14.0, 15.9]	69 [14.0, 15.9]	67 [13.5, 16.4]	47 [18.3, 06.4]	71 [15.4, 18.3]	54 [13.9, 16.7]
IR 64	83 [11.2, 19.3]	82 [11.2, 18.3]	83 [11.2, 19.3]	83 [11.2, 19.3]	78 [10.7, 17.9]	53 [15.6, 11.3]	78 [11.9, 17.9]	52 [09.2, 19.2]
Minghui 63	141 [15.4, 12.8]	136 [14.9, 12.5]	138 [15.2, 12.3]	138 [15.2, 12.3]	139 [15.3, 12.2]	96 [18.0, 07.3]	134 [16.1, 11.9]	81 [12.7, 11.1]
Pokkali	98 [09.0, 16.3]	93 [08.5, 17.2]	96 [08.8, 16.7]	94 [08.6, 17.0]	94 [08.9, 13.8]	67 [10.5, 11.9]	86 [09.1, 12.8]	71 [09.1, 18.3]
SHZ2	116 [14.3, 15.5]	113 [13.9, 15.9]	114 [14.0, 15.8]	115 [14.2, 15.7]	108 [13.6, 13.9]	83 [16.6, 13.3]	111 [15.8, 14.4]	85 [16.7, 16.5]
Sadu Cho	58 [11.6, 17.2]	54 [10.6, 18.5]	57 [11.4, 17.5]	55 [10.9, 18.2]	57 [11.4, 17.5]	41 [13.0, 7.3]	51 [16.0, 15.7]	43 [11.6, 14.0]
Swarna	131 [11.9, 16.0]	122 [11.4, 13.9]	125 [11.6, 14.4]	123 [11.4, 14.6]	121 [11.3, 14.0]	90 [14.8, 10.0]	113 [12.5, 15.9]	86 [11.8, 12.8]
Zhenshan 97	99 [13.7, 19.2]	97 [13.7, 17.5]	97 [13.7, 17.5]	97 [13.7, 17.5]	91 [13.0, 16.5]	65 [14.3, 13.8]	90 [13.7, 18.9]	56 [09.9, 23.2]
Total	1,385 [12.2, 16.0]	1,319 [11.8, 14.9]	1,346 [12.0, 15.2]	1,341 [12.0, 15.3]	1,320 [11.9, 14.2]	920 [14.7, 10.2]	1,265 [13.1, 14.6]	935 [11.9, 15.5]

Table A.6: Shown is the number of ML SNP calls broken down according to different filter criteria. We grouped the strains with respect to their population affiliation. The two performance measures [recall, FDR] are given in % in squared brackets. N/A denotes FDRs which could not be assessed as no predictions were made.

Strain	No Filter	Excluding Positions with Repetitive 25-mer Matches	Exact 25-mer Matches	Dominating 25-mer Matches	Hybridisation Quality Score ≤ 8	Sequencing Quality Score ≤ 48	Repetitive Fragments (90% Identity)	Singleton SNPs
LTH	3 [07.1, 00.0]	3 [07.1, 00.0]	3 [07.1, 00.0]	3 [07.1, 00.0]	3 [07.1, 00.0]	1 [16.7, 00.0]	3 [12.5, 00.0]	0 [00.0, N/A]
M 202	2 [03.6, 00.0]	2 [03.6, 00.0]	2 [03.6, 00.0]	2 [03.6, 00.0]	2 [03.6, 00.0]	1 [08.3, 00.0]	1 [02.4, 00.0]	1 [02.6, 00.0]
Tainung 67	0 [00.0, N/A]	0 [00.0, N/A]	0 [00.0, N/A]	0 [00.0, N/A]	0 [00.0, N/A]	0 [00.0, N/A]	0 [00.0, N/A]	0 [00.0, N/A]
Azucena	42 [14.4, 23.8]	42 [14.4, 23.8]	42 [14.4, 23.8]	42 [14.4, 23.8]	29 [10.8, 17.2]	31 [20.5, 25.8]	38 [18.2, 18.4]	37 [16.1, 27.0]
Cypress	53 [10.9, 15.1]	48 [10.0, 14.6]	53 [10.9, 15.1]	49 [10.2, 14.3]	43 [09.7, 07.0]	40 [18.0, 15.0]	51 [11.5, 13.7]	39 [12.5, 12.8]
Moroberekan	33 [10.3, 21.2]	31 [09.5, 22.6]	33 [10.3, 21.2]	31 [09.5, 22.6]	30 [09.9, 16.7]	23 [13.8, 04.3]	33 [11.3, 21.2]	22 [11.1, 18.2]
Dom Sufid	0 [00.0, N/A]	0 [00.0, N/A]	0 [00.0, N/A]	0 [00.0, N/A]	0 [00.0, N/A]	0 [00.0, N/A]	0 [00.0, N/A]	0 [00.0, N/A]
Dular	158 [19.1, 20.9]	151 [18.0, 21.9]	154 [18.5, 21.4]	154 [18.5, 21.4]	122 [15.6, 16.4]	99 [22.0, 11.1]	145 [21.6, 21.4]	106 [19.3, 20.8]
FR 13A	167 [19.8, 18.6]	164 [19.3, 18.9]	166 [19.6, 18.7]	164 [19.3, 18.9]	133 [16.6, 14.3]	111 [24.4, 13.5]	160 [21.1, 18.1]	121 [19.4, 16.5]
N 22	164 [14.9, 22.6]	160 [14.5, 22.5]	162 [14.6, 22.8]	160 [14.5, 22.5]	131 [12.3, 19.8]	106 [18.3, 12.3]	154 [15.9, 20.1]	118 [16.3, 22.0]
Rayada	221 [21.3, 13.1]	214 [20.6, 13.1]	219 [21.2, 12.8]	214 [20.6, 13.1]	178 [17.7, 10.1]	155 [27.7, 5.8]	210 [23.8, 13.3]	139 [21.9, 10.1]
Aswina	103 [21.3, 14.6]	101 [21.0, 13.9]	101 [21.0, 13.9]	101 [21.0, 13.9]	91 [19.8, 09.9]	68 [25.8, 08.8]	102 [23.1, 14.7]	79 [19.8, 19.0]
IR 64	117 [17.2, 12.0]	115 [16.9, 12.2]	117 [17.2, 12.0]	117 [17.2, 12.0]	93 [13.9, 10.8]	80 [25.5, 03.8]	108 [17.8, 11.1]	76 [14.7, 11.8]
Minghui 63	203 [22.2, 12.8]	200 [22.1, 12.0]	201 [22.2, 11.9]	200 [22.1, 12.0]	168 [18.9, 10.1]	144 [26.7, 08.3]	197 [23.4, 12.7]	123 [20.2, 07.3]
Pokkali	159 [15.2, 13.2]	151 [14.3, 13.9]	154 [14.6, 13.6]	151 [14.3, 13.9]	137 [13.4, 10.9]	110 [17.8, 09.1]	146 [15.5, 12.3]	113 [15.0, 15.0]
SHZ2	133 [17.2, 11.3]	131 [16.9, 11.5]	131 [16.9, 11.5]	131 [16.9, 11.5]	102 [13.7, 07.8]	92 [19.6, 07.6]	129 [18.9, 11.6]	93 [19.5, 10.8]
Sadu Cho	91 [18.6, 15.4]	82 [16.4, 17.1]	88 [17.9, 15.9]	83 [16.7, 16.9]	65 [13.8, 12.3]	65 [20.5, 07.7]	82 [25.4, 17.1]	68 [17.9, 16.2]
Swarna	179 [16.9, 12.8]	169 [15.9, 13.0]	174 [16.5, 12.6]	171 [16.1, 12.9]	135 [13.2, 09.6]	128 [21.8, 07.0]	156 [17.7, 13.5]	120 [17.1, 09.2]
Zhenshan 97	143 [19.0, 22.4]	143 [19.0, 22.4]	143 [19.0, 22.4]	143 [19.0, 22.4]	103 [13.9, 21.4]	96 [20.4, 16.7]	136 [19.5, 23.5]	102 [18.3, 21.6]
Total	1,971 [17.4, 16.0]	1,907 [16.8, 16.2]	1,943 [17.2, 16.0]	1,916 [16.9, 16.1]	1,565 [14.4, 12.7]	1,350 [21.7, 09.7]	1,851 [18.9, 15.8]	1,357 [17.4, 15.1]

Table A.7: Shown is the number of MB and ML SNP predictions. We grouped the strains with respect to their population affiliation. The two performance measures [recall, FDR] are given in % in squared brackets. N/A denotes FDRs which could not be assessed as no predictions were made. Note that we excluded Nipponbare both in the dideoxy SNP set and in the ML predictions.

Strain	Union MB/ML SNP Calls	Intersection MB/ML SNP Calls	SNP Calls Predicted by MB Only	SNP Calls Predicted by ML Only
Nipponbare	5,425 [N/A, N/A]	0 [N/A, N/A]	5,425 [N/A, N/A]	0 [N/A, N/A]
LTH	22,230 [66.7, 00.0]	3,011 [16.7, 00.0]	17,250 [50.0, 00.0]	1,969 [00.0, N/A]
M 202	27,963 [33.3, 00.0]	5,294 [08.3, 00.0]	16,973 [25.0, 00.0]	5,696 [00.0, N/A]
Tainung 67	12,650 [66.7, 00.0]	0 [00.0, N/A]	12,648 [66.7, 00.0]	2 [00.0, N/A]
Azucena	47,338 [23.2, 16.1]	15,248 [07.1, 20.0]	16,756 [02.7, 00.0]	15,334 [13.4, 16.7]
Cypress	60,267 [18.5, 16.7]	23,270 [09.0, 05.6]	15,758 [02.1, 33.3]	21,239 [07.4, 22.2]
Moroberekan	55,191 [12.6, 09.1]	22,013 [05.7, 00.0]	13,699 [00.6, 50.0]	19,479 [06.3, 09.1]
Dom Sufid	56,090 [04.7, 33.3]	2,029 [00.0, N/A]	53,013 [04.7, 33.3]	1,048 [00.0, N/A]
Dular	148,315 [22.0, 12.0]	63,053 [10.8, 00.0]	34,218 [02.0, 33.3]	51,044 [09.2, 17.8]
FR 13A	150,106 [26.9, 14.5]	65,741 [12.2, 02.0]	35,224 [02.8, 26.7]	49,141 [11.9, 21.7]
N 22	126,721 [20.8, 10.9]	57,792 [11.6, 06.3]	30,367 [02.8, 06.7]	38,562 [06.5, 19.5]
Rayada	151,980 [29.0, 05.6]	65,587 [14.8, 00.0]	35,043 [02.7, 12.5]	51,350 [11.6, 10.3]
Aswina	159,901 [30.4, 08.8]	66,681 [12.9, 00.0]	34,517 [05.0, 20.0]	58,703 [12.5, 11.8]
IR 64	137,257 [26.5, 07.0]	57,887 [11.9, 00.0]	33,656 [02.6, 33.3]	45,714 [11.9, 05.3]
Minghui 63	139,845 [29.8, 07.0]	60,037 [14.0, 04.2]	31,344 [03.4, 10.5]	48,464 [12.3, 09.0]
Pokkali	131,995 [18.9, 10.9]	55,493 [08.2, 06.1]	34,988 [01.8, 23.1]	41,514 [08.9, 12.3]
SHZ2	114,047 [21.9, 13.6]	47,640 [11.8, 00.0]	34,336 [03.5, 37.5]	32,071 [06.7, 17.1]
Sadu Cho	144,447 [21.2, 08.8]	60,374 [08.9, 00.0]	37,001 [03.1, 25.0]	47,072 [09.2, 10.0]
Swarna	135,179 [24.3, 07.6]	53,733 [08.4, 00.0]	38,146 [04.4, 17.2]	43,300 [11.5, 08.7]
Zhenshan 97	103,453 [24.0, 16.1]	38,328 [08.2, 11.1]	33,795 [05.1, 16.7]	31,330 [10.7, 19.2]
Total	1,930,400 [23.8, 10.3]	763,211 [10.7, 02.9]	564,157 [03.2, 21.1]	603,032 [09.9, 13.7]

Table A.8: Shown is the number of MB and ML SNP predictions at non-repetitive sites. We grouped the strains with respect to their population affiliation. The two performance measures [recall, FDR] are given in % in squared brackets. N/A denotes FDRs which could not be assessed as no predictions were made. Note that we excluded Nipponbare both in the dideoxy SNP set and in the ML predictions.

Strain	Union MB/ML SNP Calls	Intersection MB/ML SNP Calls	SNP Calls Predicted by MB Only	SNP Calls Predicted by ML Only
Nipponbare	3,968 [N/A, N/A]	0 [N/A, N/A]	3,968 [N/A, N/A]	0 [N/A, N/A]
LTH	19,546 [66.7, 00.0]	2,905 [16.7, 00.0]	14,915 [50.0, 00.0]	1,726 [00.0, N/A]
M 202	25,254 [33.3, 00.0]	5,197 [08.3, 00.0]	14,550 [25.0, 00.0]	5,507 [00.0, N/A]
Tainung 67	10,755 [66.7, 00.0]	0 [00.0, N/A]	10,753 [66.7, 00.0]	2 [00.0, N/A]
Azucena	43,812 [23.2, 16.1]	14,963 [07.1, 20.0]	14,221 [02.7, 00.0]	14,628 [13.4, 16.7]
Cypress	55,866 [18.5, 16.7]	22,795 [09.0, 05.6]	13,042 [02.1, 33.3]	20,029 [07.4, 22.2]
Moroberekan	50,881 [12.6, 09.1]	21,537 [05.7, 00.0]	11,107 [00.6, 50.0]	18,237 [06.3, 09.1]
Dom Sufid	51,817 [04.7, 33.3]	2,022 [00.0, N/A]	48,747 [04.7, 33.3]	1,048 [00.0, N/A]
Dular	140,476 [22.0, 12.0]	62,016 [10.8, 00.0]	29,703 [02.0, 33.3]	48,757 [09.2, 17.8]
FR 13A	142,760 [26.9, 14.5]	64,937 [12.2, 02.0]	30,089 [02.8, 26.7]	47,734 [11.9, 21.7]
N 22	120,363 [20.8, 10.9]	57,041 [11.6, 06.3]	26,233 [02.8, 06.7]	37,089 [06.5, 19.5]
Rayada	144,459 [29.0, 05.6]	64,697 [14.8, 00.0]	30,277 [02.7, 12.5]	49,485 [11.6, 10.3]
Aswina	151,572 [30.4, 08.8]	65,579 [12.9, 00.0]	29,834 [05.0, 20.0]	56,159 [12.5, 11.8]
IR 64	130,182 [26.5, 07.0]	57,037 [11.9, 00.0]	29,164 [02.6, 33.3]	43,981 [11.9, 05.3]
Minghui 63	133,370 [29.8, 07.0]	59,329 [14.0, 04.2]	26,783 [03.4, 10.5]	47,258 [12.3, 09.0]
Pokkali	125,361 [18.9, 10.9]	54,811 [08.2, 06.1]	30,332 [01.8, 23.1]	40,218 [08.9, 12.3]
SHZ2	108,710 [21.9, 13.6]	47,192 [11.8, 00.0]	30,215 [03.5, 37.5]	31,303 [06.7, 17.1]
Sadu Cho	136,917 [21.2, 08.8]	59,450 [08.9, 00.0]	32,370 [03.1, 25.0]	45,097 [09.2, 10.0]
Swarna	127,955 [24.3, 07.6]	52,889 [08.4, 00.0]	33,521 [04.4, 17.2]	41,545 [11.5, 08.7]
Zhenshan 97	98,833 [24.0, 16.1]	37,969 [08.2, 11.1]	30,220 [05.1, 16.7]	30,644 [10.7, 19.2]
Total	1,822,857 [23.8, 10.3]	752,366 [10.7, 02.9]	490,044 [03.2, 21.1]	580,447 [09.9, 13.7]