Detailed Materials and Methods

1. Plant materials

Twenty rice varieties were chosen for re-sequencing (Table S1). These varieties included 17 of those considered previously (1) and 3 additional nominations - Tainung 67, Minghui 63, and Zhenshan 97B. Each variety was purified by one round of single seed descent (SSD) wherein a single representative plant from each variety was chosen, its panicles were bagged at flowering, and seed collected from those panicles. SSD seed were planted in multiple pots in the IRRI screenhouse under irrigated conditions. These plants were used for collection of young leaf tissue, and at flowering, panicles were bagged with the resulting seed bulked. This generation was planted in the field for further increase of seed stocks.

2. DNA isolation

From 50 to 100 g of young leaf tissue was harvested from the plants grown in the IRRI screenhouse, frozen in liquid N_2, and ground to a fine powder using a mortar and pestle. Genomic DNAs were extracted by the CTAB/Sarkosyl extraction procedure (2) modified for use as a large scale preparation. Genomic DNAs were dissolved in TE buffer and treated with RNAse, and subjected to $CsCl_2$-ethidium bromide density gradient ultracentrifugation (3). Ethidium bromide was removed from the genomic DNAs by extraction against phenol::$CHCL_3$ (one time) followed by extraction against $CHCl_3$ (two times). Traces of phenol/$CHCl_3$ were removed by ether extraction. These DNA solutions were subjected to two rounds of precipitation against ethanol to remove $CsCl_2$. DNA pellets after the final precipitation step were dissolved in TE. DNA concentrations were estimated by Picogreen QuantIt assay (Invitrogen) followed by normalization of concentrations to 500 μg/ml.

3. Reference genome masking and target selection.

Build 4 of the rice genome sequence from the International Rice Genome Sequencing Project site (http://rgp.dna.affrc.go.jp/E/IRGSP/Build4/build4.html) sequences was used for the array design. The genome sequence was repeat masked using repeat databases from both The Institute for Genome Research(4), and the rice transposable elements database (5). Masking was done through the Integrated Transposable Element Annotation System (ITEAS) analysis pipeline developed by the Bureau Lab at McGill University (5, 6). After masking, perl scripts were used to extract the unmasked regions and save them as FASTA records. The unmasked region amounted to 240 Mb. To ensure uniqueness of the regions, each of these sequences were compared against each other using the basic local alignment tool (BLAST) (7). The sequences were then classified based on the number of hits they acquired. Those sequences classified as unique (no hits or a single hit, 91.6 Mb), and with 2 to ten hits (77.6 Mb) were chosen for oligo design.

Primers for long-range PCR amplicons spanning from 3-10 kb of the target sequence regions were designed by Oligo 6 using high quality thresholds with overlap of neighboring amplicons for continuity of coverage. Amplicons were scored and ranked by genic (coding) content, using Rice Annotation Project release 2 (RAP2) (8), and TIGR annotations (9). Each basepair in a curated RAP2 gene annotation contributed 2 points to the score of an amplicon, while the presence of a basepair in a predicted RAP2 or TIGR gene prediction contributed one point to the score. Through this ranking, 7,852 amplicons, including the 76 amplicons used for the development phase, with high genic content were selected. These amplicons covered 60.3 Mbp of nonrepetitive (unmasked) sequence, including the 345 kb of the development phase. This cut-off was selected to balance genic content of the final sequence selection with even distribution across the whole genome. An additional 5,730 amplicons with lower rankings were selected to amplify areas underrepresented in the previous step, covering 39.7 Mbp of

nonrepetitive (unmasked) sequence. The largest remaining gap (excluding large unsequenced regions) was 494 Kb.

4. Array design, sample preparation, and hybridization.

The 13,582 selected LR-PCR amplicons span 11,343 non-overlapping sequence fragments and cover 117,834,417 bp of unmasked genomic sequence, this fraction of the genome was used as the reference sequence for high-density oligonucleotide array design. Six resequencing arrays were designed such that 100,104,806 bp of the reference sequence were queried using a tiling strategy previously described (10-12) and includes short stretches of $\leq$ 60 repetitive bases. The arrays were synthesized by Affymetrix using light-directed photolithography in conjunction with chemical coupling to direct the synthesis of 25-mer oligonucleotides.

Purified genomic DNA for 20 rice strains were adjusted to $OD_{260.}$ Singleplex PCR reactions for each of the 13,586 LR-PCR amplicons were performed as follows (per reaction): 15 ng of genomic DNA from one of the 20 strains was amplified using 0.5 µM of each LR-PCR primer, 0.3 U MasterAmp extra long Taq polymerase (Epicentre Technologies), 17 µg/ml TaqStart Antibody (Clontech), 0.1 X TaqStart Antibody buffer (Clontech), 0.4 mM dNTPs, 23 mM Tricine, 3% DMSO, 45 mM Trizma, 2.4 mM MgCl2, 12.6 mM (NH4)SO4, 2.5X MasterAmp PCR Enhancer with Betaine (Epicentre Technologies), in a volume of 6 µl. The reactions were performed using a Perkin-Elmer 9700 thermocycler as follows: initial denaturation for 3 min at 95 ºC; 10 cycles at 94 ºC for 2 s and 64 ºC for 15 min; 28 cycles starting at 94 ºC for 2 s and 64 ºC for 15 min with a 20 s increase per cycle, and a final extension of 60 min at 62 ºC.

Amplicons from one strain to be hybridized together on the array were combined into one tube (~8 Mb), purified and fragmented as previously described (11). The fragmentation reactions were then labeled with either biotin or digoxigenin for 90 min

at 37 ºC in 1X One-Phor-All Buffer PLUS (Amersham), 13600 U recombinant TdT (Roche Applied Science), and either 0.1 mM each of biotin-16-ddUTP and biotin-16-dUTP (Roche Applied Science) or 0.1 mM each of digoxigenin-11-ddUTP and digoxigenin-11-dUTP (Roche Applied Science).

The arrays, each containing ~20 Mb of tiled sequence, were physically segmented into three chambers. Each chamber was hybridized with a different DNA/hybridization mixture containing labeled target DNAs at ~8 Mb complexity. The labeled target DNA was prepared for hybridization by combining biotin-labeled amplicons from one strain and digoxygenin-labeled amplicons from another strain, and then added to hybridization buffer to give a final solution of 0.1 mg/ml herring sperm DNA (Promega), 0.01 mg/ml human Cot-1 DNA (Invitrogen) 2.7 M TMACL, 56 mM MES, 5% DMSO, 2.5X Denhardt's solution, 5.8 mM EDTA, 0.01% Tween-20, 0.04 nM b-948 biotin control oligo (Sigma-Aldrich), and 0.04 nM dig-948 digoxigenin control oligo (Sigma-Aldrich). The hybridization mixture was denatured for 10 min at 99 ºC and then cooled to 50 ºC. Hybridization of the target DNA to the microarrays took place at 52 ºC for 18-20 h with constant rotation.

After hybridization, the arrays were rinsed with 6X SSPE and stained for detection of the biotin- and digoxigenin-labeled hybridized targets by 20 min incubations at room temperature using the following series of 4 stain reagents (each of which was in 6X SSPE, 1X Denhardt's solution, and 0.01% Tween-20): Stain 1 with 2.5 µg/ml Anti-Digoxigenin Ab, clone 1.71.256, mouse IgG1 (Roche Applied Science) plus 5 µg/ml streptavidin (Invitrogen); Stain 2 with 5 µg/ml anti-streptavidin (Rabbit) Biotin Conjugated (Rockland Immunochemicals); Stain 3 with 5 µg/ml Alexa Fluor 647–R phycoerythrin goat anti-mouse IgG (Invitrogen) plus 1 µg/ml streptavidin Alexa Fluor 488 conjugate; and Stain 4 with 9 µg/ml AffiniPure Mouse Anti-Goat IgG (H+ L). To enhance the digoxygenin signal, the arrays were then incubated further with

Stain 3 followed by Stain 4. The arrays were rinsed with 6X SSPE, 0.01% Tween-20 between incubations, and washed at high stringency in 0.2X SSPE, 0.01% Tween-20 for 60 min at 37 ºC after the completion of staining. After rinsing with 1X MES, the arrays were then scanned using custom-built confocal scanners.

5. Base-calling and SNP detection.

We used the same pattern recognition (model-based or MB) algorithms for analysis of the resequencing data that had previously been described for human and mouse genomic resequencing and SNP discovery (11, 12). For this project, a strict base call is made for a sequence position when the ratio of the brightest to next-brightest feature is greater than a threshold of 1.3 for biotin-labeled DNA and 1.1 for digoxigenin-labeled DNA, and the conformance around that position is at least 0.80 on both strands. A relaxed base call is made if these criteria are met for just one strand and the other strand is ambiguous (that is, it did not pass either the intensity ratio or conformance requirements). For alternate base calls that do not match the reference sequence, we also require that there are no brighter alternate calls meeting these criteria within +5 or -5 bases.

6. Trace files and quality scores.

Trace files for each contiguous fragment of tiled sequence for each of the strand orientations were created and quality scores for each assigned base were determined using algorithms previously described (11). These trace files have been uploaded to the NCBI Trace File Archive at (http://www.ncbi.nlm.nih.gov/Traces/).

7. Annotation of Repetitive Probe Sets in the Reference Genome

The special design of SNP detection arrays, with eight probes for each site in the target DNA, increases the chance of cross-hybridization and therefore unspecific signals. Cross-hybridization of repetitive sequences can either suppress a true SNP signal and thus reduce sensitivity, or generate an incorrect SNP signal, decreasing specificity. For rice, approximately three fourths of the genome is repetitive, which further contributes to the problem of cross-hybridization. Because repetitive sequences were not completely excluded from the ~100 Mb interrogated with the arrays, it was important to identify oligomers occurring more than once, such that repetitive sites could be treated separately in subsequent analyses.

Repetitive probes were annotated by identifying oligomers that match at least to one other 25-mer in the target DNA, allowing for some degree of degeneracy. Because all four possible nucleotides were represented at the central position of each 25-mer, we allowed mismatches at the center base in addition to peripheral mismatches. The same mismatch criteria were used as in the *A. thaliana* resequencing project (10), distinguishing between the three match types *exact*, *inexact* and *short 25-mer matches* (see (10), Supporting Online Material, p. S4/5). Additionally, we included bulged mismatches. Bulges in hybridizing oligomers are formed when one or more nucleotides remain unpaired (13). The definition of *bulged 25-mer matches* is restricted to a one-base bulge that is located only on one of the strands (14). Bulged 25-mer matches are then matches between an oligomer of length 25 and an oligomer of length 26 in which the longer one of the pairing strands contains a bulge of exactly one nucleotide, tolerating mismatches at the center position. The first and last positions were not included in this analysis as they are not real bulges, but dangling ends, which are covered by the definition of short 25-mer matches. Note that the definitions of the four match types are exclusive, i.e. the described sets are disjoint.

We generated a list of all 25-mers contained in both the probe DNA, i.e. the DNA that was used for tiling and immobilization on the arrays, and the amplified target reference. 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 certain mismatches. Finally, the sorted list was linearly traversed to report all 25-mer occurring more than once, restricted to matches only between tiled and target DNA 25-mers. We processed the bulged 25-mer data such that only match pairs with distance of at least 25 bp in the genome were included in statistics and further analyses. This filter strategy was motivated by the observation that poly(N)-regions in the reference genome sequence contribute heavily to the number of bulged 25-mer matches.

In total, 5,160,864 positions were annotated as repetitive, making up 5.16% of all positions used for tiling. False positives are most likely to be observed, and of consequence, at positions where the counts of the nucleotide at the center position of the repetitive 25-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. Information on both the repetitive and dominating 25-mer positions was used in the machine learning algorithm described later.

8. Dideoxy resequencing for quality assessment.

Supervised learning methods – as used for SNP discovery from the resequencing data – require labeled examples both for training and error evaluation. The *A. thaliana* resequencing project benefited from a data set previously sampled by dideoxy sequencing (15). For the rice resequencing project, such a curated data set of polymorphisms was not previously available. We therefore used dideoxy sequencing of a series of fragments to enable compilation of a set of true SNP and non-SNP positions.

A subset of the Perlegen tiled regions were chosen for dideoxy-based resequencing by randomly selecting fragments to ensure a representative evaluation (14). For each candidate resequencing region, the program Primer3 was used to pick primer pairs that would amplify products averaging 600 bp (16), and the standard M13 forward and reverse primer sequences were added to the ends of the sequence-specific primers. A total of 1,440 PCR products were amplified from a randomly assigned cultivar. An additional 1,536 PCR products were each amplified from three randomly assigned cultivars. A total of 6,048 PCR products were sequenced with M13 forward and reverse primers on an ABI 3730. Raw sequences were trimmed for quality (17) and assembled by phrap (18). ClustalW (19) 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 an unrealistically high (more than 50) number of SNPs. After the preprocessing step, the dideoxy sequencing set comprised 1,755,395 positions across all varieties of which 9,499 positions were assigned to be polymorphic. A summary of the number of SNPs detected by dideoxy sequencing and their distribution over the different genomic regions is given in a table at http://www.OryzaSNP.org.

9. Experimental inputs for the Machine Learning algorithm

The hybridization intensities, denoted by *I*, were mean fluorescence measurements for each of the four bases A, C, G and T on each of the forward and reverse strand. Furthermore, a raw base call *B* referred to the base for which the hybridization intensity was highest within a probe quartet. Additionally, quality scores *QS,* estimating the error probability of calling a certain base, were used.

Normalization of hybridization data allowed for the correction of between-array variation and to obtain comparability of the data generated by multiple array experiments. The hybridization data were quantile-normalized (20) on the level of amplicon pools across all varieties (14). Because the quantile-normalization algorithm required identical batch sizes, but the amplicon pools did not all contain the same number of probes, each pool was filled up with sham intensities to the identical maximal pool size. The sham intensities were sampled from the observed distributions. We excluded data from pool 0 on wafer 0 (developmental array), as the number of probes in this pool was 13-fold lower than in the other 15 pools. The intensities from pool 0 were normalized by averaging the outcomes when sorting according to the normalized intensities from five pools randomly selected from the 15 other pools.

10. A Machine Learning (ML) method for SNP identification

We applied a two-layered approach based on Support Vector Machines (SVMs) (21, 22) to predict SNPs from the hybridization data similar to the approach used in the *A. thaliana* resequencing project (10). In a first step, SVMs were trained using information comprising the array data, sequence characteristics and repetitiveness of the genome based on the results of the annotation of repetitive 25-mers. Since the hybridization had been normalized beforehand, we were able to train machines across all varieties instead of using a separate machine for each variety as was done for *A. thaliana*. After whole genome SNP predictions had been made independently for each variety in the first step, we trained a second layer of SVMs, which were able to integrate information across varieties, as they were provided with results from layer 1 for all varieties as input. Specifically, we re-examined all positions for which in at least one other variety a SNP had been predicted with the layer 1 SVM. 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 that had 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 varieties. Finally, each prediction was assigned a confidence value reflecting the posterior likelihood of a true SNP prediction. To train the SVMs and to evaluate the performance of the classifier, the data set of known polymorphic sites was utilized.

a. Layer 1 SVM:

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 could be discarded using the following filter criteria, and this resulted in a significantly smaller data set, which also reduced computational time in both training and prediction. Specifically, we excluded positions that were identical to the reference with high probability as well as positions where the corresponding array data gave inconsistent information on the called base.

The first criterion was met for a given position *p* in the target variety *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 hybridization 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* that passed the filter as the set:

$$P = \{p_t \mid t \text{ a given target variety}\}$$
$$\text{where } p_t = \{p \mid B_t^+(p) = B_t^-(p) \neq RS(p) \wedge B_{ref}^+(p) = B_{ref}^-(p) = RS(p) \wedge \overline{QS}_t(p) > 5\}$$

Input generation. To be able to train SVMs on the given data set, we had to generate an input vector $\mathbf{x}_p^1$ for each position *p* that had passed the filter in layer 1. Each $\mathbf{x}_p^1$ includes measurements at this position and neighboring positions within ±4 bp around *p*: Both maximal intensities I_{max} of each of two quartets and averages of the non-maximal intensities I_{sec} of each of the quartets at each position in the 9 bp window were included. As the machine learning method was found to be most accurate for SNPs separated by 7 to 30 bp compared to the model based approach on the *A. thaliana* resequencing data, we chose a window size of 9 bp to obtain predictions that were complementary to a high degree to the SNP calls made by the model based method.

Moreover, we added ratios of the maximal intensities at *p* and its neighboring positions (Q_1) and ratios of the maximal intensities at *p* of the target and the reference variety (Q_2). Using these quotients as input features was motivated by the shape of a typical SNP signal reflected in the intensity pattern. $\mathbf{x}_p^1$ 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 repeat analysis to include occurrence counts *k* of repetitive 25-mers at *p*.

As we used the same SVM predictor for all varieties, we additionally included information on the variety of which the intensities were taken, denoted as *v*. Table S2 describes all inputs in detail.

The input vectors were normalized on the training set to mean 0 and standard deviation 1 per input dimension. Additionally, each input example was normalized using $\frac{\mathbf{x}_p^1}{||\mathbf{x}_p^1||}$ for all positions *p*. The normalized input vectors were then used to train SVMs with an RBF kernel (21) using the SHOGUN toolbox (http://www.shogun-toolbox.org/) (23), which allows a fast and efficient training.

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 permuted order, and is also important for model selection. We randomly split our data set of labeled positions into five equally-sized disjoint subsets $S_1,\ldots,S_5$ with respect to a uniform distribution of positives per variety. 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 that was used for model selection was then defined as $X_i = \{S_j \mid j \neq i\}$. Thus, the set X_i at iteration step i consisted of 80% of all labeled 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}.

To find the best model, we applied a measurement based on 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 ($1-\text{specificity}$). The higher the AUC value is, the more accurate the classifier. In our case, we only wanted to maximize the area limited to a maximal ratio of true positives to false positives (i.e., 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. The area a_{ikj} between this curve and the line showing one false positive at five true positives was then calculated corresponding to FDRs below $\frac{1}{6}$. 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.

Prediction and output transformation. The prediction for each position that passed the filter was made 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 for prediction.

As we used outputs from five different SVMs, the predictions were not directly comparable. 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.

This was done by means of estimating a monotonic piecewise linear transformation 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 robust 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 $x(q) \le V \le x(q+1)$ true positives, i.e. the number of known SNP positions, with prediction values *V* in the range and *n(q)* the number of all labeled 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 (24) was used.

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

$$c = \begin{cases} y(1) & V \le x(1) \\ \dfrac{y(q+1)\cdot(V - x(q)) + y(q)\cdot(x(q+1) - V)}{x(q+1) - x(q)} & x(q) \le V \le x(q+1) \\ y(40) & V \ge x(40) \end{cases}$$

Details of layer 1 SVMs are given in Table S3.

b. Layer 2 SVM

Filter criteria. In the second step, we exploited information from layer 1 predictions across all varieties: Only positions with confidence values c_t greater than some threshold th_t in at least one other variety t were used for training the SVMs of layer 2. This threshold was determined per variety across all test sets 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 labeled positions in *t* that passed the filter. We further used a relaxed filter for layer 2 to take positions into account that were likely polymorphic in at least one variety, but which did not have raw base calls identical on both strands in

the variety 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 positions with a mean hybridization quality score $\overline{QS}_t(p)$ less than or equal to 5.

All positions *P* that passed filter 2 are described as:

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

where [[ ]] denotes the indicator function.

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 varieties were included. To be able to connect a position to its target variety *t*, we encoded the variety information by vector of length $19^2 = 361$ with *b* at the 19 positions corresponding to variety *t* and zeros elsewhere. Confidence values were encoded in the same way. The variety information *v* of a position was omitted. As an additional feature, we included information on polymorphisms that distinguish the *ssp. indica* variety 93-11 (*ind*) from the reference genome sequence of the *ssp. japonica* variety Nipponbare, which was used in the array design. This further facilitated the detection of SNPs at known polymorphic positions. We aligned the two genome sequences using MUMMER (http://mummer.sourceforge.net), a suffix-tree algorithm for large-scale genome

alignments (25) and parsed SNPs from the alignments with more than 90% identity within a 61 bp window. See Table S3 for more details on the layer 2 inputs.

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

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 used these trained machines to make predictions for the tiled regions of the entire genome, including unlabelled positions that were interrogated on the hybridization resequencing arrays. As theses positions had neither been employed either for training nor evaluation, any of the five SVMs could be used for prediction.

For all of the 19 non-reference varieties, a layer 1 SVM was chosen at random to make a prediction for each unlabelled position that passed filter 1. We also made predictions for positions across all varieties that met the filter criteria in at least one other variety. 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 that 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.

c. Base calling

As the output of the machine learning method only comprised the probability $C(p)$ for being a true SNP at a given position p, the corresponding observed base $B(p)$ was inferred from the hybridization data by applying the following base calling algorithm:

- CASE 1

if $B^{+}(p) \neq B^{-}(p) \wedge B^{+}(p) \neq RS(p) \wedge B^{-}(p) \neq RS(p)$

set $B(p) = \mathrm{N}$

- CASE 2

if $B^{+}(p) = B^{-}(p) = RS(p)$

set $B(p) = RS(p)$

- CASE 3

if $B^{+}(p) = B^{-}(p) \neq RS(p) \wedge C(p) \geq 0.855$

set $B(p) = B^{+/-}(p)$

- CASE 4

if $B^{+}(p) = B^{-}(p) \neq RS(p) \wedge C(p) < 0.855$

set $B(p) = \mathrm{N}$

- CASE 5

if $B^{+}(p) = RS(p) \wedge B^{-}(p) \neq RS(p) \wedge C(p) \geq 0.855$

set $B(p) = B^{-}(p)$

- CASE 6

if $B^+(p) = RS(p) \wedge B^-(p) \neq RS(p) \wedge C(p) < 0.855$

set $B(p) = \mathrm{N}$

- CASE 7

if $B^-(p) = RS(p) \wedge B^+(p) \neq RS(p) \wedge C(p) \geq 0.855$

set $B(p) = B^+(p)$

- CASE 8

if $B^-(p) = RS(p) \wedge B^+(p) \neq RS(p) \wedge C(p) < 0.855$

set $B(p) = \mathrm{N}$

d. Performance evaluation

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 that are recovered by the predictor, is denoted by recall or sensitivity:

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

Repetitive positions and those with low sequencing quality were used for training, but were excluded from performance evaluation (14). The results are shown in Table S2.

11. SNP annotation.

All SNPs were located relative to the IRGSP (6) and TIGR (9) pseudomolecules. Because the IRGSP and TIGR pseudomolecules are not identical, a conservative approach was taken for locating SNPs on the TIGR pseudomolecules. All of the Perlegen PCR amplicons were located relative to the TIGR pseudomolecules by using the program Vmatch (http://www.vmatch.de/) to place the 5' and 3' amplicon primers on the pseudomolecules. Only alignments to the expected pseudomolecule or the TIGR unanchored contigs were allowed, and the alignments had to produce an appropriately sized region ± 100 bp. A total of 13,558 out of 13,582 Perlegen amplicons were successfully mapped to the TIGR pseudomolecules. Tiled regions were also aligned to the TIGR pseudomolecules using Vmatch. Because of sequence differences and notable gap differences between the IRGSP and TIGR pseudomolecule sequences, a small number of tiled regions were mapped using Gmap (26). Tiled region alignments were only accepted when they occurred on the expected chromosome or on the TIGR unanchored contigs pseudomolecule. Additionally, tiled regions were required to align within Perlegen PCR amplicons that had been mapped on the TIGR pseudomolecules. A total of 54,998 out of 55,019 tiled regions were successfully mapped to the TIGR pseudomolecules. SNPs locations were mapped to the TIGR pseudomolecules by using Vmatch to align 101 bp segments surrounding each SNP. Only alignments that resided within mapped tiled regions and on the expected chromosome or the unanchored contigs pseudomolecule were accepted. For the MBML-intersect SNP set, a total of 158,928 out of 159,879 IRGSP localized SNPs were successfully mapped to the TIGR pseudomolecules.

Annotation of all SNPs was performed relative to the both the IRGSP and TIGR pseudomolecules and to both the RAP and TIGR gene models (8, 27). A Chado-schema, generic model organism database (28) was used to store SNP, gene model and pseudomolecule data. Custom Perl scripts and database queries were designed to classify all SNPs based on whether they reside within genes or within intergenic regions; these data are viewable in the browser. For those SNPs that were located within TIGR gene models, the SNPs were further classified as to their position within the gene: 5'-UTR, 3'-UTR, intron or CDS. SNPs that were positioned within coding sequence were analyzed for the effect that they had on the relevant codon, and non-synonymous and synonymous mutations were noted. Finally, major effect SNPs were also annotated if they were located within intron donor or acceptor sites, created a premature stop codon, destroyed a stop codon or destroyed an initiation codon (Table S4 and summarized queries at http://www.OryzaSNP.org). All annotations were stored within the database.

To examine the distribution of non-synonymous and synonymous SNPs within the rice genome, several analyses were performed. All TIGR rice gene models were aligned to Arabidopsis gene models using BLASTP (7). The best alignments with p-values better than 10^{-5} were used to classify rice genes into gene families that have also been identified within Arabidopsis (29). Pfam protein domains were identified within the TIGR rice gene translations using iprscan (9, 30, 31). Paralogous rice genes that are found within segmentally or tandemly duplicated regions of the rice genome have been previously identified (9, 27, 32). The number of non-synonymous and synonymous SNPs were identified that occurred within genes with gene family classifications, within Pfam domains, within tandemly or segmentally duplicated genes or within genes annotated as expressed/unexpressed and known/unknown (summarized at http://www.OryzaSNP.org).

SNP sites annotated as non-synonymous, synonymous or as large-effect changes were extracted from the MBML-intersect data set, and only sites with high confidence base calls for at least 15/20 varieties were included in calculations of allele frequencies. For each SNP site, the number of varieites with bases different from the reference were counted and plotted by frequency and annotation category.

12. Validation of large effect SNPs.

Sixty loci containing SNPs annotated as large effect were randomly selected for SNP validation (additional information at http://www.OryzaSNP.org). Custom scripts were written to automatically generate primers that could be used to amplify a region of DNA that contained major effect SNPs that had been targeted for resequencing. A boulder formatted file containing 800 bp regions surrounding each targeted SNP was created and used as input to the program Primer3 (16). Each primer pair was designed to amplify a product between 550 and 650 bp. The M13 forward (TGTAAAACGACGGCCAGT) and reverse (CAGGAAACAGCTATGACC) primer sequences were added to the 5' ends of each primer pair in order to allow the use of common primers during sequencing.

PCR was performed on all cultivars predicted to contain a large effect SNP at a given locus, and PCR products were cleaned using the Exo-sap procedure as described by the manufacturer (USB Corporation, Cleveland, OH). Dideoxy sequencing was performed at 2x coverage for all amplicons as described above.

Reference genome sequences containing 500 bp up and downstream of putative SNP sites were extracted from the OryzaSNP database. Reference sequences and trace files were entered into Lasergene, Seqman software (33), and contigs were assembled

using a mismatch criteria of 80%. SNP reports were generated for position 500 for all contigs and nucleotide mismatches were validated by manually inspecting the trace files (Table S4).

13. Summary statistics and dendrogram construction.

Summary statistics the MBML-intersect were calculated by chromosome using scripts under Microsoft Excel. The numbers of total SNP sites as well as the breakdown to those at repetitive sites (as ascertained from the probe-by-probe analysis) and non-repetitive sites are reported. The total number of clear genotypes across all 20 varieties is given. The frequencies of allele classes (ignoring ambiguous calls) were calculated for those with a frequency of 0.05 (singletons), with a frequency of 0.1 (two varieties have the allele), and frequencies of >0.1 (three or more varieties carry the allele). The number of sites with bi-allelic data and their breakdown into transitions and transversions is reported. Information on the data sets is available at http://www.OryzaSNP.org.

Pairwise distance matrices using the simple matching coefficient were calculated for the SNPs at non-repetitive sites the MBML-interscet datasets by using scripts under Microsoft Excel. SNPs designated as "N" were treated as missing. An unweighted Neighbor Joining tree was constructed using DARwin 5 (http://darwin.cirad.fr/darwin, (34) (Figure 2a). Trees for the other datasets and tables of un-ambiguous SNPs occurring between pairs of varieties are available at http://www.OryzaSNP.org.

14. Linkage Disequilibrium analysis.

For the analysis of LD, only biallelic non-singleton SNPs in the MBML-intersect dataset were considered. We calculated the LD as the correlation coefficient r^2 between SNP pairs. The mean r^2 value was calculated for bins of size 10 kb based on all pairs of

non-singleton SNPs. Due to the extensive population structure in the sample of 20 varieties, we examined LD decay in each subpopulation separately. Because of the small sample size in the aus group, only the indica and japonica groups, with eight varieties in each group, were analyzed. Only SNP pairs with no missing data at both loci in at least six chromosomes of the eight varieties are included in the calculations (Fig. 2c).

15. Introgression analysis.

To study the component of ancestral groups along the genome in each variety, especially those regions introgressed from other subgroups, we applied a likelihood ratio test method. Based on the three groups, indica (8 varieties), japonica (8 varieties) and aus (4 varieties), obtained from population structure analysis, we looked at all of the putative introgressions between the pairs of groups. Those SNPs in the MBML-intersect dataset with considerable missing data were removed. In addition, SNPs not segregating in the varieties and those with less than three, three and two genotyped varieties in the indica, japonica and aus groups, respectively, were also excluded from the analysis.

Using the introgression between indica and japonica as an example, for any given window of 100 Kb with at least 10 SNPs, the ratio of the average sharing of each variety with the group it belongs to and the other group is calculated. When calculating the average sharing, at least three pairs of comparison in each group were required. If the average sharing ratio is less than 0.5, it was defined as an introgression region. Introgressed blocks for each accession are shown in Fig. 3 and SI Fig. 1. If the source of an introgression is ambiguous, both potential donors are indicated. For each 100 kb window across the genome, the maximum occurrence of the introgression types (indica, japonica or aus with ambiguity allowed) is plotted at the top of each chromosome.

For each variety and type of introgression, the length and number of 100 kb windows are calculated. The number of introgressions is calculated as the total of noncontiguuos blocks of the same type. Then, the average length in Mbp of an introgression is tallied. For each variety, the number of introgression blocks of the same type shared with other varieties is also given (Table S6).

16. Haplotype sharing.

The extent of haplotype sharing among accessions was examined by splitting the genome into non-overlapping 100 kb windows and calculating the proportion of differences between all pairs of accessions in each window. Only SNPs with calls in both members of the pair being examined were included. All runs of more than five consecutive 100 kb windows with fewer than 10% difference between pairs of accessions were identified. When a 100 kb window had less than 5 SNP comparisons for a pair, this window was not counted towards the minimum of five windows, but were allowed to extend a run. The resulting blocks of SNP similarity between all pairs of accessions are shown in SI Fig. 2 for chromosomes 1-12.

References

1. McNally, KL*, et al.* (2006) Sequencing multiple and diverse rice varieties. Connecting whole-genome variation with phenotypes. *Plant Physiol* 141**:**26-31.
2. Fulton, TM, Chunwongse, J, Tanksley, SD (1995) Microprep protocol for extraction of DNA from tomato and other herbaceous plants. *Plant Mol Biol Rep* 13**:**207-209.
3. Maniatis, T, Fritsch, EF, Sambrook, J (1982) (Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.).
4. Yuan, Q*, et al.* (2003) The TIGR rice genome annotation resource: annotating the rice genome and creating resources for plant biologists. *Nucleic Acids Res* 31**:**229-233.
5. Juretic, N, Bureau, TE, Bruskiewich, RM (2004) Transposable element annotation of the rice genome. *Bioinformatics* 20**:**155-160.

6. IRGSC (2005) The map-based sequence of the rice genome. *Nature* 436**:**793-800.
7. Altschul, SF, Gish, W, Miller, W, Myers, EW, Lipman, DJ (1990) Basic local alignment search tool. *J Mol Biol* 215**:**403-410.
8. Rice Annotation Project (2008) The Rice Annotation Project Database (RAP-DB): 2008 update. *Nucleic Acids Res* 36**:**D1028-D1033.
9. Ouyang, S*, et al.* (2007) The TIGR Rice Genome Annotation Resource: improvements and new features. *Nucleic Acids Res* 35**:**D883-887.
10. Clark, RM*, et al.* (2007) Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. *Science* 317**:**338-342.
11. Frazer, KA*, et al.* (2007) A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. *Nature* 448**:**1050-1053.
12. Patil, N*, et al.* (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* 294**:**1719-1723.
13. Karaman, MW, Groshen, S, Lee, CC, Pike, BL, Hacia, JG (2005) Comparisons of substitution, insertion and deletion probes for resequencing and mutational analysis using oligonucleotide microarrays. *Nucleic Acids Res* 33**:**-.
14. Bohnert, R (2007) (Friedrich Miescher Labor of the Max Planck Society and Eberhard Karls Universität, Tübingen).
15. Nordborg, M*, et al.* (2005) The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biol* 3**:**e196.
16. Rozen, S , Skaletsky, H (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* 132**:**365-386.
17. Chou, HH , Holmes, MH (2001) DNA sequence quality trimming and vector removal. *Bioinformatics* 17**:**1093-1104.
18. Green, P (1996) Documentation for Phrap. http://bozeman.mbt.washington.edu, Genome Center, University of Washington.
19. Thompson, JD, Higgins, DG, Gibson, TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22**:**4673-4680.
20. Bolstad, BM, Irizarry, RA, Astrand, M, Speed, TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19**:**185-193.
21. Schölkopf, B , Smola, AJ (2002) *Learning with Kernels* (MIT Press, Cambridge).
22. Vapnik, VN (1982) *Estimation of dependences based on empirical data* (Springer, New York).
23. Sonnenburg, S, Rätsch, G, Schäfer, C, Schölkopf, B (2006) Large scale multiple kernel learning. *J Mach Learn Res* 7**:**1531–1565.
24. Sonnenburg, S, Schweikert, G, Philips, P, Behr, J, Ratsch, G (2007) Accurate splice site prediction using support vector machines. *BMC Bioinformatics* 8 Suppl 10**:**S7.
25. Kurtz, S*, et al.* (2004) Versatile and open software for comparing large genomes. *Genome Biol* 5**:**R12.
26. Wu, TD , Watanabe, CK (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* 21**:**1859-1875.

27. Campbell, MA, *et al.* (2007) Identification and characterization of lineage-specific genes within the Poaceae. *Plant Physiol* 145**:**1311-1322.
28. Mungall, CJ , Emmert, DB (2007) A Chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinformatics* 23**:**i337-346.
29. Rhee, SY, *et al.* (2003) The *Arabidopsis* Information Resource (TAIR): a model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. *Nucleic Acids Res* 31**:**224-228.
30. Finn, RD, *et al.* (2006) Pfam: clans, web tools and services. *Nucleic Acids Res* 34**:**D247-251.
31. Zdobnov, EM , Apweiler, R (2001) InterProScan--an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* 17**:**847-848.
32. Lin, H, Zhu, W, Silva, JC, Gu, X, Buell, CR (2006) Intron gain and loss in segmentally duplicated genes in rice. *Genome Biol* 7**:**R41.
33. Kleywegt, GJ (1992-2005) (Uppsala University, Uppsala, Sweden), p. SEQMAN.
34. Perrier, X, Jacquemoud-Collet, JP (2006), p. DARwin software http://darwin.cirad.fr/darwin