ARTICLE

A partial regression coefficient analysis framework to infer candidate genes potentially causal to traits in recombinant inbred lines

Jan Michael Yap[1]*, Ramil Mauleon[2], Eduardo Mendoza[1,3], and Henry Adorna[1]

[1]Department of Computer Science, University of the Philippines Diliman, Quezon City Philippines
[2]T.T. Chang Genetic Resources Center, International Rice Research Institute, Los Baños, Laguna, Philippines
[3]Department of Membrane Biochemistry, Max Planck Institute of Biochemistry, Martinsried, Germany

An important task in genetics research is to be able to generate a set of candidate genes causal towards expression of a particular trait. In this paper, a partial regression coefficient analysis framework to infer candidate genes causal to a trait, particularly for recombinant inbred lines, is proposed. The framework integrates marker data, quantitative trait loci data, phenotypic data, and gene expression and location data to generate the set of candidate genes. The framework was applied to infer candidate causal genes to salt tolerance in rice. Results suggest 91 genes to be potentially causal to salt tolerance in rice. Review of the literature substantiated the results in mentioning the role of genes related to cell wall and membrane-related function, signal transduction, and gene expression-related functions to salt tolerance in rice.

INTRODUCTION

An important utility of computational sciences in biology, apart from building models of biological organisms and systems, is to generate hypotheses, which could then be tested later using wet laboratory experiments. In genetics, a crucial hypothesis generation task is to identify candidate genes potentially causal to the expression of a phenotypic trait. This is essential to narrowing down genes that most likely constitute the genetic architecture of certain traits. Traditional genetic studies would usually rely on one type of data set to be able to infer a set of candidate genes. There is a fairly recent shift though from relying on one specific type of data for candidate gene set generation to integrating several types of data to be able to perform the aforementioned task.

Integrative genomics and inferring genes causal to phenotypic traits

Integrative genomics (Schadt et al. 2005), or sometimes termed as systems genetics (Druka et al. 2008, Lusis et al. 2008), has steadily gained ground over the past couple of years as a means of conducting genetics and genomics research. Integrative genomics, in a nutshell, aims to integrate several types of genetic and genomic data sets such as gene expression, sequence, and phenotypic data using mathematical, statistical, and computational techniques. The goal of integrative genomics is to gain a more or less holistic view of the underlying mechanisms of a

*Corresponding author
Email Address: jcyap@dcs.upd.edu.ph
Submitted: October 21, 2013
Revised: March 8, 2014
Accepted: March 20, 2014
Published: June 25, 2014
Editor-in-charge: Eduardo A. Padlan
Reviewers: Jingky P. Lozano-Kühne and Ricardo C.H. del Rosario

KEYWORDS
partial regression coefficient analysis, recombinant inbred lines, genetic markers, gene expression, quantitative trait loci, candidate causal genes

biological process that cannot be obtained via analysis of only one type of data. It has been utilized, among others, to build and fine tune genetic network models (Chu et al. 2009, Druka et al. 2008), to improve the resolution of genetic mapping (Degnan et al. 2008), and to find associations between genetic and molecular functions with certain phenotypic traits (Al-Shahrour et al. 2005).

An interesting application of integrative genomics is inferring which genes are causal towards the expression of a phenotypic trait. Inferring candidate genes using integrated genetic and genomic data sets provides a multi-faceted approach to the analysis leading to better inference and even the elucidation of finer details not apparent when only one kind of data is used (Lusis et al. 2008, Schadt et al. 2005). The typical strategy for such analysis is the integration of sequence data, gene expression data, and phenotypic data (Kang et al. 2012, Schadt et al. 2005).

One of the earliest works involving integrative genomics-based gene inference was done by (Schadt et al. 2005) to infer genes causal towards mice obesity. The work involved integration of genetic marker data, phenotypic data, and microarray gene expression data via a Bayesian framework. The method hinged on the premise that proximity of quantitative trait loci (QTLs) of mice obesity and QTLs to gene expression traits (termed expression QTLs or eQTLs) is considered evidence that the expression of a gene causes the expression of a trait.

The same framework was used in a later work in identifying candidate genes for type 2 diabetes (Kang et al. 2012). Genome wide association studies (GWAS) were leveraged to generate sequence related data sets in this work. It should be noted that as GWAS was used, the researchers thus had disease-specific single -nucleotide polymorphisms (SNPs) as the kind of sequence data used in the analysis. This is the difference from the previous work of Schadt et al. (2005) wherein QTLs were used.

Recombinant inbred lines

In genetics and genomics, and especially in plant biology (Druka et al., 2008, Thomson et al. 2010, Walia et al. 2005), recombinant inbred lines (RILs) are very popular subjects for study. RILs are organisms that were bred by mating two inbred strains, then repeatedly performing sibling cross or selfing (Broman 2005). The resulting genome of such organisms can be described as a mosaic of the original genomes of the parents.

One of the reasons RILs are popular is that the genomic structure of a RIL is preserved (Broman 2005). This means that replicates of a RIL would have the same genotype as the original RIL. This reduces the cost, time, and resource needed since genotyping on the replicates is no longer necessary, given that once the genotype of a RIL is known, then the replicates would then share the same genotype as that of the original.

Another advantage of using RILs is that the strains almost always have homozygous alleles, *i.e*., the allele at a genetic locus is inherited purely from one of the parents (Collard et al. 2005). This is especially a favorable property in QTL mapping, wherein the locus/loci causing significant changes to the expression of a trait is/are being identified (Broman 2005, Collard et al. 2005). The idea is that a QTL is a locus where the mean trait values of individuals having a particular allele at the QTL differ significantly from those of individuals having another allele. Hence, the basic idea of QTL mapping is that given a locus, individuals are grouped together based on the allele at that particular locus, and the mean trait value of each group is thus computed and evaluated. Some problems arise when the allele at a locus is heterozygous, *i.e*., the locus allele is inherited in part from one parent, and in part from the other. RILs therefore provide a clear delineation from which parent did the RIL inherit the allele at a particular locus, and would thus provide a relatively easier means to perform QTL mapping.

Sufficiency of gene expression dataset samples to integrative genomics framework

There is an abundance of publicly accessible gene expression data sets from several repositories online. However, there is scarcity of data sets with sufficient samples for analysis of very specific traits, *e.g*., salt tolerance in rice. For instance, the rice salt stress gene expression data set from Walia et al. (2005) contains microarray gene expression data from 11 individuals. However, the individuals only come from 2 cultivars: FL478 and IR29, both of which are RILs. If one were to apply the framework used in (Schadt et al. 2005) to infer genes causal to rice salt tolerance, then the number of individuals that can be used for analysis is effectively cut down to only 2. Thus, this would make the data set unsuitable for further analysis using the given framework. One possible workaround would be to generate more gene expression data particularly for individuals not covered by the available data set, although this might prove costly[1]. A challenge would then be to infer candidate causal genes given the available data without resorting to generating additional gene expression data.

Partial regression coefficient framework for inferring candidate genes potentially causal to traits in RILs

A partial regression coefficient analysis framework is proposed to infer candidate causal gene expression to phenotypic traits on RILs. The proposed framework integrates marker and QTL data, phenotypic data, and gene expression and location data to generate the set of candidate genes. The highlight of the proposed framework is that the inference of candidate genes does not rely heavily on gene expression data, but instead on the

[1]Price per array for rice from Affymetrix as of December 2008 is US$400. In rice data set scenario, this is multiplied by 30 arrays needed for *de facto* bare minimum in statistical analyses, total cost would be around US$12,000

inferred allelic state of the genes based on the states of the nearest markers. This particular property of the framework avoids the pitfall of having the number of individuals for the whole analysis be based on the effective number of individuals used in gene expression analysis, as mentioned earlier. The number of individuals will now rely on the sample size of the marker and QTL data sets. It should be noted that gene expression values would still be used although as part of a filtering step, which will be shown in a later section. Figure 1 shows the flowchart of the steps done in the proposed framework

The proposed framework has a premise similar to that of QTL mapping techniques: that the states of particular loci of the genome affect the expression of a trait. The difference is that in QTL mapping, the loci are generic locations within the genome, while in the proposed framework, the loci are necessarily coordinates of the genes. It should be noted, though, that the proposed scheme analyzes the effect of the state of gene sequences on top of the state of trait QTLs. The effect of QTL on a trait is thus still considered in the analysis and that the effect depending on the state of the genes is treated as an additional effect towards the expression of a trait.

The proposed framework thus assumes that QTL and gene expression, through its inferred allelic state, have a causal relationship to phenotypic trait. Borrowing the notations from (Schadt et al. 2005) and (Lusis et al. 2008), the underlying model of the framework is L → C ← R, where L is the QTL, R is the gene expression, and C is the phenotypic trait. The framework, however, does not aim to establish a causal relationship (or lack thereof) between QTL and gene expression, as was done in (Schadt et al. 2005).

Interpolating the allelic state of a gene

The first step in the framework, assuming all pertinent data sets were already gathered, is to interpolate the allelic state of each gene, per individual. While the allelic state of the genetic markers are known, that may not be the case for a gene. To this end, the value of the variable representing the allelic state of a particular gene is thus interpolated employing the genetic markers used in QTL mapping. This is done by first identifying the markers near the 5' and 3' ends of a gene's sequence. Using these markers and the data on their allelic states, the allelic state at each end of a gene sequence, and subsequently the whole gene itself, can thus be interpolated. A simple two-point probability formula for selfed RILs (Broman 2005) can be used to infer the allelic state of one end of the gene sequence based on the score of the nearest marker. For example, in 2-way selfed RILs, the two-point probability is computed as follows: given two points/loci in the genome of a RIL, x_i and x_j, the probability that the allelic score/state of the two points/loci are different is given by:

$$P\left(g_{x_i} \neq g_{x_j}\right) = \frac{2r}{1+2r}$$

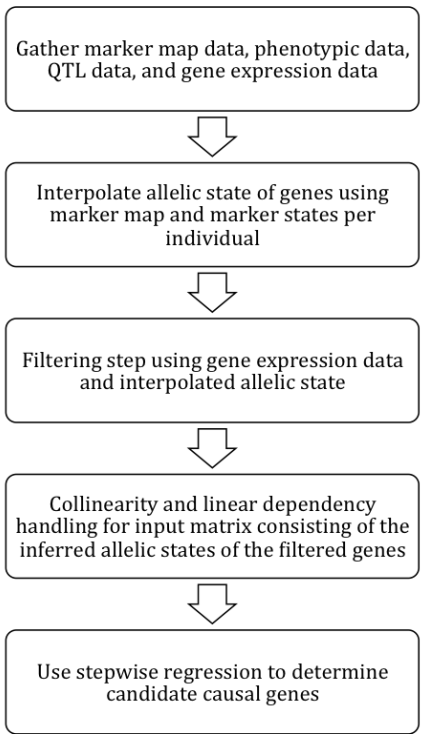

Figure 1. The steps in the proposed partial regression coefficient analysis framework.

where g_{x_i} and g_j are the are the (target) allelic states of points/loci x_i and x_j, respectively, and r is the recombination rate of x_i and x_j.

If $P\left(g_x^{i} \neq g_x^{j}\right)$ is greater than 0.5, then the allelic state of the 5' or 3' end of the gene sequence should be different from the score of the nearest marker. Otherwise, the 5' or 3' end's allelic state is the same as the nearest marker's score. To finally infer the allelic state of the gene sequence, the inferred allelic states of both 5' and 3' ends are checked. If the allelic states of both ends are the same, then the probe set assumes that allelic state. If the allelic states of both ends are different, then the gene can be thought of as having a heterozygous allelic state. Such genes are not included in the analysis, as with QTL analysis in RILs.

Filtering step using gene expression data and interpolated allelic state of the genes

As a means of data reduction, a filtering step is applied using the gene expression values and inferred allelic states as discriminants. The main idea in this filtering step is to look for genes whose (mean) expression values have statistically significant differences when segregated according to the inferred allelic state of their respective sequences. An outline of how this is done is as follows:

1. For each gene, expression values are grouped according to the inferred allelic states of the sample for that gene.
2. A two-sample mean test (*e.g.*, Student's t-test) can be

performed on the two groups to check if there is a statistically significant difference on the expression value of the gene subject to a difference in its inferred allelic state.

3. A gene is most likely affected by allelic state change if there is a significant difference between the means of the two groups, as per the statistical test.
4. If there is only one group after segregation (*i.e.*, only one state inferred for the gene for all samples), a gene can still be considered most likely affected by allelic state change if there is *no* significant difference between the mean expression values of each RIL.

The last item is actually a weaker notion of the premise of the filtering step in that a gene whose allelic state is the same across all samples implies that there is no significant change in the expression values across samples. It might seem counterintuitive or even contrary to the purpose of this step, but it is meant as a measure to maximize usage of the data set. This is especially true if there is a significantly few number of individuals included in the gene expression data set, and will be evident in the discussion of the application of the framework to rice salt stress data. The genes, with their inferred allelic states per sample, whose expression values were found to be significantly affected by changes in allelic state, comprise those which are to be subjected to partial regression coefficient analysis.

Linear model for determining candidate causal genes

The basis model for determining causal gene expression is the linear model:

$$Y_k = \mu + \varrho_k + \sum_{m=1}^{g} \beta_m z_{mk}$$

where Y_k is the value of the phenotypic trait for individual k, μ is the intercept (*e.g.*, midparent, or mean value) for the phenotypic trait, ϱ_k is the effect of QTL/s for the phenotypic trait, β_m is the contributory effect of the mth gene, and z_{mk} is the coded variable representing allelic state of the mth gene in individual k. q is the number of QTLs included in the model, while g is the number of candidate causal genes.

In the model, ϱ_k is treated as a constant and is always included during the gene selection/regression analysis step. Since a trait may have a number of QTLs, ϱ_k can thus represent the aggregated effects of a number of QTLs, and be expressed as a summation:

$$\varrho_k = \sum_{j=1}^{q} \alpha_j x_{jk}$$

where α_j is the additive effect of the jth QTL for the phenotypic trait and x_{jk} is the coded variable representing the allelic state of the jth QTL in individual k. The values for α_j and x_{jk} can be obtained/derived from prior QTL mapping endeavors, as will be shown later in the implementation. The value of q (*i.e.*, the number of QTLs included in the model) depends on the focus of the analysis. This implies that the analysis may include all the known QTLs for the trait, or may include just a subset of the known QTLs.

Dealing with multicollinearity in the inferred allelic states of genes

Incorporating the genes as regressors by using their inferred allelic states as the representative variables causes the resulting linear model to become susceptible to what is termed as the dummy variable trap. This is the condition wherein multicollinearity, *i.e.*, several variables having significant linear dependence, is present and thus may significantly bloat the standard errors of the regression model and, in some cases, would hinder the solvability of the linear model (Suits 1957). Multicollinearity may hold in the inferred allelic state genes especially when genetic linkage is present or when two or more genes are located proximal to each other. The two cases would mean that such genes would share the same inferred allelic state in most, if not all, samples, as those genes would be near the same set of markers used in inferring their allelic states. In the context of selecting candidate genes using the linear model, this might mean that the selection among those genes with (almost) similar states across samples might be done arbitrarily since they would have almost similar effects. Hence, there is a need to handle multicollinearity within the inferred allelic states.

Handling multicollinearity in the framework is a 2-phase process: the first phase is by creating clusters of highly correlated variables and the second is by removing linearly dependent variables. In the first phase, a correlation coefficient/statistic is computed between each gene using their inferred allelic states across samples. A cluster can then be established by grouping together certain genes such that pairwise correlation values of any two genes within the group exceed (or are at least equal to) a certain threshold. A gene from the cluster is used as a representative for the whole group. The choice of the representative gene may be done arbitrarily if the correlation coefficient threshold is high enough (*e.g.*, > 0.99) or by choosing the gene whose set of inferred allelic states across samples is nearest to the centroid of the cluster. The idea now is that if the representative gene is selected or eliminated in the succeeding phase and in the gene selection step, the same will be done to the other members of the cluster to which the representative gene belongs. It is expected that some genes will not be included in any cluster, but nonetheless are directly used as input to the succeeding steps of the framework together with the cluster representatives. The result of the first phase would be a reduced set of genes where genes with highly correlated allelic states across samples are collapsed into one representative gene.

The second phase would be to eliminate linear dependency from the reduced set of genes. This can be done via linear algebraic methods such as the Gaussian elimination and the Gauss-Jordan reduction methods. The main idea is to create a matrix wherein the columns consist of the inferred allelic state of genes across samples. Afterwards, an attempt is made to convert the matrix into the identity matrix, or as close to it as possible, using row operations. A linearly dependent column, which represents a gene, is one whose entries do not consist of almost entirely 0's with a solitary 1. Such columns are thus eliminated from consideration in the gene selection phase detailed in the next section.

Partial regression coefficient analysis

Now that the allelic states of the genes have been inferred, genes filtered, and multicollinearity handled, partial regression coefficient analysis can proceed. As mentioned in the previous section, QTL effects are held constant and the phenotypic trait is thus regressed on the genes using the genes' inferred allelic states. For the analysis, regression-based subset selection is done to select the genes to be included in the set of candidate genes.

There are several modes of regression-based subset selection: the forward selection method, the backward elimination method, Efroymson's algorithm, subquential replacement, and exhaustive search (Miller 2002). The forward method starts with a model with no regressors and adds a variable to the model that best fits the model. The backward method starts with a model consisting of all regressors and then eliminates one variable at a time. The Efroymson's algorithm is a bidirectional method with two versions: one is using forward selection with possible elimination of a variable already included in the model done at each step, and another is using backward elimination with possible inclusion of a variable not present in the model done at each step. Sequential replacement performs selection in the same manner as the forward version of Efroymson's algorithm, but performs replacement of variables instead of removing them. Exhaustive search essentially involves generating all possible subsets of regressors and finding the subset that gives the best-fit model.

The number of regressors to be considered in the selection will affect the computational demand of the search. Hence and in the context of the framework, the choice of regression mode would depend on the number of genes from which the candidate causal genes will be selected. Given a significantly large number of genes for consideration, the forward selection, the forward version of Efroymson's algorithm, or sequential replacement may be employed. If there is a fairly small number of genes, then backward elimination, the backward version of the Efroymson's algorithm, or exhaustive search may be used.

Determining the candidate genes potentially causal to a trait

The main goal in the partial regression coefficient analysis step is to check which genes should be considered candidate causal to the trait by checking which linear model best fits the data at hand. Since regression-based subset selection is used in the framework to do this, finding the best-fit model is tantamount to determining how many variables/genes should ultimately comprise the linear model.

To determine the appropriate number of genes, penalized statistics such as information criteria (*e.g.*, BIC, AIC) and adjusted scores (*e.g.*, adjusted R^2, Mallows Cp) computed at each step of the regression-based subset selection are taken note. Apart from checking a model's goodness of fit, a model is penalized based on the number of variables it contains. This particular feature thus avoids model over-fitting by exacting a heavy penalty on those models with significantly large number of variables in spite of the goodness of fit.

The appropriate number of genes is thus the number of genes included in the model whose penalized statistic value is an extremum (maximum or minimum, depending on the statistic used). Furthermore, the genes contained in that model comprise the set of candidate genes potentially causal to a trait. Note that a representative gene from a cluster of correlated genes determined prior may be included in this set, in which case the other members of that cluster are also treated as candidate causal genes.

MATERIALS AND METHODS

Rice salt stress markers, phenotype, and QTL

The framework was applied to rice salt stress response data to check the efficacy of the method. Marker and phenotypic data on salt tolerance were acquired from the Supplemental Data of (Thomson et al. 2010). 140 Pokkali x IR29 RILs constitute the aforementioned data set. Additional raw trait data were also requested from the authors of the aforementioned literature. Shoot sodium-potassium (Na-K) ratio was used as the representative trait for salt tolerance. The QTLs for shoot Na-K ratio identified in the article were also used in the analysis. There were 2 QTLs identified for shoot Na-K ratio: qSNK1 and qSNK9. qSNK1 is located in chromosome 1 with peak marker location at 41.2 Mb, while qSNK9 is located in chromosome 9 with peak marker at 52.8 Mb. In the analysis, the peak markers were used as the representative locations for both QTLs.

Rice salt stress gene expression

Rice salt expression microarray data were obtained via the NCBI Gene Expression Omnibus (GEO) site, under series entry reference ID GSE3053. The data set consists of 57,381 probe sets and has a total of 11 samples: 5 control and 6 salt-stressed. Two cultivars make up the data set: the salt-sensitive IR29 and the salt-tolerant FL478. The data set being microarray data, probe sets are thus used as representatives for the genes that will make up the list of candidates in the implementation of the framework.

A previous work applied differential expression analysis on the data set, of which part of the results were used in the framework's implementation (Walia et al. 2005). In the aforementioned endeavor, 958 probe sets were found to be significantly differentially expressed. Significantly differentially expressed probe sets were those whose expression values changed at least 2 -fold, at $\alpha = 0.05$, between the control and the salt-stressed samples. Additionally, the significantly differentially expressed probe sets were also characterized and grouped based on whether they are up- or down-regulated, and whether the up- or down-regulation was observed in the FL478 or IR29 cultivar. Of the 958 probe sets, 620 were up-regulated (448 in IR29, 164 in FL478, 8 in both cultivars), while 338 were down-regulated (182 in IR29, 162 in FL478, 2 in both cultivars). This grouping was maintained in the succeeding steps of the framework's implementation.

Assigning allelic states and effects of the QTLs

To designate the allelic states of the Na-K ratio QTLs, the score of the peak markers was used. Allelic states were coded as either 0 (IR29) or 1 (Pokkali). This is to be consistent with the observation of Thomson et al. (2010) that the QTLs have "increased tolerance effect from Pokkali". Hence, Pokkali markers are assumed to have a contribution to whatever additive effect there is towards the phenotype, while IR29 markers are assumed to contribute nothing. QTL effects are represented by the additive effect/s of the QTL/s multiplied by the corresponding coded allelic score/s for an individual. As mentioned in the description of the framework, the QTLs (via their additive effects and allelic scores) were treated as constants in the model, *i.e.*, they were always present in the model formed by regression analysis at each step. The additive effect of qSNK1 is 0.32 while that of qSNK9 is 0.31 (Thomson et al. 2010).

Assigning allelic states of the probe sets

Allelic states of the probe sets were inferred via the process mentioned in a previous section on gene allelic state interpolation of the framework. The 2-way RIL formula from Broman (2005) as mentioned in the similar section was used to compute the allelic state probabilities. Marker locations were provided with the obtained marker data in the Supplementary Data of Thomson et al. (2010). Probe set locations were obtained from the Supplemental Data of Walia et al. (2005). No probe sets had heterozygous inferred allelic states, hence all probe sets were included in the next step of the framework.

Filtering for genes significantly affected by inferred allelic state change

For the filtering step, only significantly differentially expressed probe sets identified in Walia et al. (2005) were included in the analysis as a means for data reduction. In the filtering step, only the salt-stressed FL478 and IR29 samples were used to check the (significance of) differences in expression values between the cultivars under salt stress, instead of under normal conditions. Thus, the analysis was further focused on the salt-stress response. The two-sample t-test of means was employed for the analysis, using $\alpha = 0.05$ as the threshold.

There were cases in the processing of the data set wherein all samples were included in only one group because the inferred allele of those genes in all samples is from IR29. The workaround in this scenario is, given a gene, to divide the samples based on their RIL/cultivar, *i.e.*, if the sample is IR29 or FL478. Afterwards, t-test was performed between the mean expression values of the two cultivar-based groups and then the computed p -value was checked if it is *above* the α threshold, *i.e.*, there is no significant difference in the mean expression values between the two cultivars. Probe sets that satisfy the given condition were thus included.

After the filtering step, 33 FL478 down-regulated, 29 FL478 up-regulated, 37 IR29 down-regulated, and 87 IR29 up-regulated probe sets were identified for inclusion in the next step of the framework. Tables 1-4 contain the list of identified probe set IDs in the filtering step, together with the inferred allelic state per cultivar, and the corresponding t-test p-values.

Forming the clusters of correlated genes

Pearson's correlation was performed on the inferred allelic states to create the correlation matrix. The correlation coefficient value threshold used in creating the clusters was 1. This means that those probe sets contained in a cluster have the same inferred allelic states across all samples, considering that the allelic state is a binary variable (0 or 1). Table 5 presents the clusters formed with 2 or more elements. Note that only the representative probe set per cluster will be used and hence there was a reduced number of probe sets as regressors. 22 FL478 down-regulated, 24 FL478 up-regulated, 27 IR29 down-regulated, and 43 IR29 up-regulated probe sets were used as input in the method used for regression-based subset selection.

Removing linear dependency and partial regression coefficient analysis

The *regsubsets* function of the *leaps* R package (Lumley 2009) was utilized for the regression analysis. On top of its regression capabilities, the function also removes linearly dependent columns in the matrix of regressors prior to performing regression and hence performs the second phase of the multicollinearity-handling step. Additionally, *regsubsets* also performs variable reordering to ensure that the inclusion of variables in the linear model at each step is not dependent on the sequence by which they were analyzed. Considering the reduced number of probe sets used as regressors, exhaustive search was chosen as the mode of subset-selection regression used in the experiment.

Three linear model variations were used in the analysis: one, where 2 shoot Na-K ratio QTLs were included, and two, where

Table 1. FL478 down-regulated probe sets identified in the filtering step to be those whose expression values are affected by allelic state change.

Probe Set ID	Inferred Allelic State		t-test p-value
	IR29 Cultivar	FL478 Cultivar	
Os.34782.1.S1_at	0	1	0.048220049
OsAffx.21835.1.S1_x_at	0	1	0.048278309
Os.46481.3.S1_at	0	1	0.024080793
OsAffx.13586.2.S1_at	0	0	0.504924175
Os.29814.1.S1_at	0	0	0.612129626
OsAffx.25967.2.S1_at	0	0	0.514388775
OsAffx.25968.1.S1_at	0	0	0.764776992
OsAffx.15286.1.S1_at	0	0	0.765241418
OsAffx.32296.1.A1_at	0	0	0.207319058
OsAffx.15397.1.S1_x_at	0	0	0.194018833
Os.53094.1.S1_at	0	0	0.136979391
Os.17112.1.S1_at	0	0	0.561178322
OsAffx.7216.1.S1_at	0	0	0.382948358
OsAffx.15760.1.S1_at	0	0	0.741147719
Os.6345.1.S1_at	0	0	0.140163674
Os.1503.1.S1_at	0	0	0.349865497
OsAffx.5111.1.S1_x_at	0	0	0.429477841
Os.23216.2.S1_x_at	0	0	0.706845127
Os.7449.2.S1_x_at	0	0	0.468800225
OsAffx.16888.1.S1_at	0	0	0.484777346
OsAffx.5782.1.S1_at	0	0	0.625206568
OsAffx.8550.1.S1_at	0	0	0.862743685
OsAffx.18135.1.S1_at	0	0	0.894490707
OsAffx.22071.1.S1_at	0	0	0.11318583
OsAffx.30032.2.S1_at	0	0	0.177235457
Os.10765.1.S1_at	0	0	0.758092816
OsAffx.30986.1.S1_at	0	0	0.068943352
OsAffx.31017.1.S1_at	0	0	0.107246971
Os.5354.1.S1_at	0	0	0.396482824
OsAffx.19547.1.S1_at	0	0	0.856316082
Os.57108.1.S1_at	0	0	0.464299402
Os.51000.1.S1_at	0	0	0.115116004
OsAffx.19880.1.A1_at	0	0	0.882548696

For the second and third columns, 0 means the inferred allele of the gene is from IR29, while 1 means the inferred

Table 2. FL478 up-regulated probe sets identified in the filtering step to be those whose expression values are affected by allelic state change.

Probe Set ID	Inferred Allelic State		t-test p-value
	IR29 Cultivar	FL478 Cultivar	
Os.57128.1.S1_at	0	0	0.770772149
OsAffx.27871.1.S1_at	0	0	0.331682546
OsAffx.15353.1.S1_at	0	0	0.484170594
OsAffx.7652.1.S1_at	0	0	0.658950405
OsAffx.17017.1.S1_at	0	0	0.163855492
Os.4291.1.S1_at	0	0	0.053467683
OsAffx.7539.1.S1_at	0	0	0.32870849
OsAffx.25844.1.S1_at	0	0	0.291752435
OsAffx.30227.1.S1_at	0	0	0.331875218
OsAffx.18144.1.S1_at	0	0	0.453407364
OsAffx.29212.1.S1_at	0	0	0.129092325
OsAffx.15672.1.S1_at	0	0	0.98102065
OsAffx.7116.1.A1_at	0	0	0.317104819
OsAffx.28239.1.S1_at	0	0	0.099215222
OsAffx.30860.1.S1_at	0	0	0.815084779
OsAffx.15322.1.S1_x_at	0	0	0.256169484
OsAffx.25794.1.S1_at	0	0	0.313581107
Os.39552.1.A1_s_at	0	0	0.609705215
OsAffx.20772.1.S1_at	0	0	0.482043826
OsAffx.30282.1.S1_at	0	0	0.156963763
OsAffx.448.1.S1_at	0	0	0.232068028
OsAffx.4165.1.S1_at	0	0	0.566484262
Os.18257.1.S1_at	0	0	0.242854193
OsAffx.5351.1.S1_at	0	0	0.182443535
OsAffx.1477.1.A1_at	0	0	0.902707673
Os.51641.1.S1_x_at	0	0	0.069184218
OsAffx.25987.2.A1_at	0	0	0.581916752
Os.50556.2.S1_at	0	0	0.351108871
OsAffx.30030.2.S1_at	0	1	0.028848379

For the second and third columns, 0 means the inferred allele of the gene is from IR29, while 1 means the inferred allele of the gene is from Pokkali.

Table 3. IR29 down-regulated probe sets identified in the filtering step to be those whose expression values are affected by allelic state change.

Probe Set ID	Inferred Allelic State		t-test p-value
	IR29 Cultivar	FL478 Cultivar	
OsAffx.15084.1.S1_at	0	1	0.014385517
Os.53892.1.S1_at	0	1	0.031444633
OsAffx.31005.1.S1_at	0	0	0.398196464
OsAffx.24637.2.S1_x_at	0	0	0.247596911
OsAffx.15917.1.S1_at	0	0	0.051376154
OsAffx.5780.1.A1_at	0	0	0.483256756
OsAffx.13705.1.S1_at	0	0	0.066775643
OsAffx.4989.1.S1_at	0	0	0.319412789
OsAffx.25807.1.S1_at	0	0	0.700190395
OsAffx.5047.1.S1_at	0	0	0.89923966
OsAffx.8723.1.S1_at	0	0	0.415466989
OsAffx.9449.1.S1_at	0	0	0.520945327
Os.40417.1.A1_at	0	0	0.814044764
OsAffx.24436.1.S1_at	0	0	0.715538157
OsAffx.29285.1.S1_at	0	0	0.270987802
OsAffx.5019.1.S1_at	0	0	0.981152996
OsAffx.31859.1.S1_at	0	0	0.713489157
Os.23264.1.A1_at	0	0	0.523076135
OsAffx.29742.2.S1_x_at	0	0	0.448601362
OsAffx.15417.1.A1_at	0	0	0.813113113
Os.15521.2.S1_at	0	0	0.078847728
OsAffx.28377.1.S1_at	0	0	0.22492827
OsAffx.8455.1.S1_at	0	0	0.700168202
OsAffx.25958.1.S1_at	0	0	0.195298756
OsAffx.30902.1.S1_at	0	0	0.895322758
OsAffx.31890.1.S1_at	0	0	0.325716079
OsAffx.18890.1.S1_at	0	0	0.625635603
OsAffx.19518.1.S1_at	0	0	0.993708278
Os.12750.1.S1_x_at	0	0	0.3727254
Os.52357.1.S1_at	0	0	0.887413344
Os.8082.1.S1_at	0	0	0.471736654
OsAffx.4155.1.S1_at	0	0	0.809219186
OsAffx.4851.1.S1_x_at	0	0	0.447940315
OsAffx.4918.1.S1_at	0	0	0.588004384
Os.47942.1.A1_at	0	0	0.395965119
OsAffx.31668.1.S1_at	0	0	0.153503646
OsAffx.19862.1.S1_at	0	0	0.461571583

For the second and third columns, 0 means the inferred allele of the gene is from IR29, while 1 means the inferred allele of the gene is from Pokkali.

Table 4. IR29 up-regulated probe sets identified in the filtering step to be those whose expression values are affected by allelic state change.

Probe Set ID	Inferred Allelic State		t-test p-value IR29 Cultivar
	IR29 Cultivar	FL478 Cultivar	
Os.53710.1.S1_at	0	1	0.0144259
Os.2558.1.S1_a_at	0	1	0.007896554
Os.35288.1.S1_at	0	1	0.001736393
Os.17491.1.S1_at	0	1	0.022198848
Os.51602.1.S1_at	0	1	0.023252334
Os.27143.1.S1_at	0	1	0.02358489
Os.52577.1.S1_x_at	0	1	0.001374542
Os.11330.1.S2_at	0	1	0.017373274
Os.8413.2.A1_at	0	1	0.015308532
Os.320.2.S1_a_at	0	1	0.018416527
Os.8413.2.A1_x_at	0	1	0.023903438
Os.49855.1.S1_at	0	1	0.04316882
Os.8413.2.A1_a_at	0	1	0.010786621
Os.28427.1.S1_x_at	0	1	0.035514625
Os.11469.1.S1_at	0	1	0.008253926
Os.24865.1.A1_at	0	1	0.016274039
Os.456.3.S1_s_at	0	1	0.015402546
Os.49997.1.S1_at	0	1	0.049816598
Os.2957.1.S1_at	0	1	0.042830264
Os.52577.1.S1_at	0	1	0.011078716
Os.26486.1.S1_at	0	1	0.02848959
Os.52211.1.S1_at	0	1	0.034406095
Os.50024.1.S1_at	0	1	0.009045332
OsAffx.25340.1.S1_at	0	1	0.004521062
OsAffx.26914.1.S1_at	0	1	0.007987203
Os.16903.1.A1_at	0	1	0.015957801
Os.24471.1.S1_at	0	1	0.029015162
Os.37729.1.S1_s_at	0	1	0.025754135
OsAffx.7680.1.S1_at	0	0	0.211706677
OsAffx.13585.1.S1_at	0	0	0.792685661
Os.55786.1.S1_at	0	0	0.93413792
Os.17906.1.S1_at	0	0	0.060829831
OsAffx.27821.1.S1_at	0	0	0.774454345
Os.46053.2.S1_x_at	0	0	0.129304619
OsAffx.29090.1.S1_at	0	0	0.520664883
OsAffx.13758.1.S1_at	0	0	0.789505027
OsAffx.5120.1.S1_x_at	0	0	0.402419114

continued on the next page

continuation of Table 4

Probe Set ID	Inferred Allelic State		t-test p-value IR29 Cultivar
	IR29 Cultivar	FL478 Cultivar	
OsAffx.7131.1.S1_at	0	0	0.565901963
Os.20817.5.A1_at	0	0	0.618089277
Os.51533.1.S1_at	0	0	0.062292644
OsAffx.16866.1.S1_at	0	0	0.568947835
Os.17419.1.S1_at	0	0	0.058752545
OsAffx.5111.1.S1_x_at	0	0	0.429477841
OsAffx.25822.1.S1_at	0	0	0.345043512
OsAffx.27824.1.S1_at	0	0	0.965373727
Os.37062.2.S1_at	0	0	0.46554832
Os.12415.1.S1_at	0	0	0.775053478
Os.53553.1.S1_at	0	0	0.148861367
OsAffx.4896.1.S1_at	0	0	0.361435214
Os.14214.1.S1_at	0	0	0.328052615
OsAffx.29191.1.S1_at	0	0	0.157593172
Os.46555.1.S1_at	0	0	0.716216226
Os.27247.1.S1_at	0	0	0.059931061
Os.8098.1.S1_at	0	0	0.055104057
OsAffx.18881.1.S1_at	0	0	0.549868171
Os.2371.1.S1_at	0	0	0.050342281
OsAffx.10020.1.S1_x_at	0	0	0.51149931
Os.55116.1.S1_at	0	0	0.079045802
OsAffx.19559.1.S1_at	0	0	0.454767547
Os.49830.1.S1_at	0	0	0.116671523
Os.31468.2.S1_at	0	0	0.848824815
Os.21957.1.S1_at	0	0	0.241743187
OsAffx.27747.1.S1_at	0	0	0.090710245
Os.48441.1.S1_at	0	0	0.116492623
OsAffx.450.1.S1_at	0	0	0.738349843
Os.15058.2.S1_at	0	0	0.770736849
OsAffx.5283.1.S1_at	0	0	0.084739515
Os.51775.1.S1_x_at	0	0	0.883892866
Os.25570.1.S1_at	0	0	0.433604145
Os.21635.1.S1_at	0	0	0.434529589
OsAffx.5122.1.S1_at	0	0	0.150233345
Os.5265.1.S1_x_at	0	0	0.271687021
Os.8511.1.S1_s_at	0	0	0.178389816
Os.4449.2.S1_at	0	0	0.559763445
Os.27205.1.S1_at	0	0	0.219334317

continued on the next page

continuation of Table 4

Probe Set ID	Inferred Allelic State		t-test p-value IR29 Cultivar
	IR29 Cultivar	FL478 Cultivar	
OsAffx.16161.1.S1_x_at	0	0	0.591168512
OsAffx.31987.1.S1_x_at	0	0	0.569121725
Os.38806.1.A1_x_at	0	0	0.193035871
OsAffx.27513.1.S1_s_at	0	0	0.830153116
OsAffx.15765.1.S1_at	0	0	0.706640435
Os.27155.1.S1_at	0	0	0.100786574
OsAffx.15535.1.S1_at	0	0	0.218314203
OsAffx.25790.1.S1_at	0	0	0.209051657
Os.27790.1.A1_at	0	0	0.28299876
OsAffx.15364.1.S1_at	0	0	0.231010952
Os.4631.1.S1_at	0	0	0.166231296
Os.8956.1.S1_at	0	0	0.70638654

For the second and third columns, 0 means the inferred allele of the gene is from IR29, while 1 means the inferred allele of the gene is from Pokkali.

only one of the QTLs was included. The idea behind varying the included QTLs was to cover as much ground as possible, as inclusion of some probe sets may be dependent on the configuration of QTLs included in the linear model.

Regarding the choice of penalized statistics, it should be first noted that there are several measures of goodness of fit computed by the *regsubsets* function: BIC, adjusted R^2, and Mallows Cp. In the implementation, the Mallows Cp statistic was used to determine the best-fit model with the appropriate number of genes. The measure was chosen because the behavior of the plot of the values presents a definitive extremum, or in this case minimum, which the plot of the values of the other two statistical measures did not show. Mallows Cp thus allows for decisive determination on the appropriate number of genes for this experiment. Figure 2 shows the Mallows Cp plots obtained in the experiment.

RESULTS

91 probe sets were inferred potentially causal to salt tolerance based on the causality modeling result. Of the 91 probe sets, 20 were from the FL478 down-regulated set, 12 were from the FL478 up-regulated set, 17 were from the IR29-down regulated set, 41 were from the IR29 up-regulated set, and 1 probe set (OsAffx.5111.1.S1_x_at) in both FL478 down-regulated and IR29 up-regulated sets. In the analysis where the 2 QTLs are included in the model, 13 probe sets were included from the FL478 down-regulated set (Mallows Cp = 1.4878057), 10 probe sets from the FL478 up-regulated set (Mallows Cp = -3.014615), 13 probe sets from the IR29 down-regulated set (Mallows Cp = -3.544527), and 38 probe sets from the IR29 up-regulated set (Mallows Cp = -14.11369). In the case where only qSNK1 was included in the model, 17 FL478 down-regulated genes (Mallows Cp = 4.168814), 9 FL478 up-regulated genes (Mallows Cp = 2.659841), 14 IR29 down-regulated genes (Mallows Cp = -3.194516), and 38 IR29 up-regulated genes (Mallows Cp = -12.53976) were included in the model. In the model where only qSNK9 is included, 15 FL478 down-regulated genes (Mallows Cp = 2.544221), 10 FL478 up-regulated genes (Mallows Cp = 21.23064), 4 IR29 down-regulated genes (Mallows Cp = 21.9653), and 24 IR29 up-regulated genes (Mallows Cp = 13.47789) were determined as the number of genes for inclusion.

There were genes that were included in multiple scenarios according to which of the QTLs were included in the regression model. For example, the FL478 down-regulated probe set with Affymetrix ID OsAffx.13586.2.S1_at, FL478 up-regulated probe set OsAffx.27871.1.S1_at, IR29 down-regulated OsAffx.4155.1.S1_at, and IR29 up-regulated Os.52577.1.S1_x_at were included in all three scenarios. Other examples of genes included in multiple scenarios are FL478 up-regulated OsAffx.30227.1.S1_at (model with both qSNK1 and qSNK9 included, and with qSNK9 only), IR29 down-regulated OsAffx.24637.2.S1_x_at (model with both qSNK1 and qSNK9 included, and with qSNK1 only), and IR29 up-regulated Os.20817.5.A1_at (model with qSNK1 only and with qSNK9 only included). Tables 6-9 give the summary of the inclusion of the causal genes based on the QTLs included in the model.

Table 5. Clusters formed after the first phase of multicollinearity handling.

	Clusters
FL478 Down	Os.34782.1.S1_at*, OsAffx.21835.1.S1_x_at
	OsAffx.13586.2.S1_at*, Os.29814.1.S1_at
	OsAffx.25967.2.S1_at*, OsAffx.25968.1.S1_at
	OsAffx.15286.1.S1_at*, OsAffx.32296.1.A1_at
	OsAffx.15397.1.S1_x_at*, Os.53094.1.S1_at
	OsAffx.15397.1.S1_x_at*, Os.17112.1.S1_at
	OsAffx.7216.1.S1_at*, OsAffx.15760.1.S1_at, Os.6345.1.S1_at
	Os.1503.1.S1_at*, OsAffx.5111.1.S1_x_at
	OsAffx.16888.1.S1_at*, OsAffx.5782.1.S1_at
	OsAffx.30986.1.S1_at*, OsAffx.31017.1.S1_at
	Os.34782.1.S1_at*, OsAffx.21835.1.S1_x_at
FL478 Up	OsAffx.27871.1.S1_at*, OsAffx.15672.1.S1_at
	OsAffx.15353.1.S1_at*, Os.4291.1.S1_at, OsAffx.15322.1.S1_x_at
	OsAffx.17017.1.S1_at*, OsAffx.29212.1.S1_at
	OsAffx.25844.1.S1_at*, OsAffx.25987.2.A1_at
IR29 Down	OsAffx.31005.1.S1_at*, OsAffx.18890.1.S1_at
	OsAffx.24637.2.S1_x_at*, OsAffx.8723.1.S1_at, OsAffx.15417.1.A1_at, OsAffx.4851.1.S1_x_at, OsAffx.4918.1.S1_at
	OsAffx.15917.1.S1_at*, OsAffx.5047.1.S1_at
	OsAffx.13705.1.S1_at*, OsAffx.25958.1.S1_at, Os.8082.1.S1_at
	OsAffx.4989.1.S1_at*, OsAffx.5019.1.S1_at
	OsAffx.25807.1.S1_at*, Os.47942.1.A1_at
IR29 Up	Os.2558.1.S1_a_at*, Os.17491.1.S1_at
	Os.52577.1.S1_x_at*, Os.8413.2.A1_at, Os.8413.2.A1_x_at, Os.8413.2.A1_a_at, Os.11469.1.S1_at, Os.52577.1.S1_at
	Os.11330.1.S2_at*, Os.2957.1.S1_at
	Os.28427.1.S1_x_at*, Os.456.3.S1_s_at, Os.49997.1.S1_at, Os.50024.1.S1_at, OsAffx.26914.1.S1_at
	Os.24865.1.A1_at*, Os.16903.1.A1_at
	OsAffx.13585.1.S1_at*, OsAffx.25790.1.S1_at
	Os.55786.1.S1_at*, Os.14214.1.S1_at, Os.31468.2.S1_at, Os.5265.1.S1_x_at, Os.4449.2.S1_at, Os.38806.1.A1_x_at, OsAffx.27513.1.S1_s_at
	OsAffx.27821.1.S1_at*, OsAffx.4896.1.S1_at, OsAffx.27747.1.S1_at, Os.27155.1.S1_at, OsAffx.15535.1.S1_at, Os.27790.1.A1_at, OsAffx.15364.1.S1_at, Os.4631.1.S1_at
	OsAffx.29090.1.S1_at*, OsAffx.16866.1.S1_at, Os.49830.1.S1_at
	OsAffx.13758.1.S1_at*, OsAffx.25822.1.S1_at, Os.8511.1.S1_s_at
	OsAffx.5120.1.S1_x_at*, OsAffx.5111.1.S1_x_at, Os.27247.1.S1_at, Os.55116.1.S1_at, OsAffx.5122.1.S1_at, Os.27205.1.S1_at
	OsAffx.7131.1.S1_at*, Os.8098.1.S1_at, OsAffx.18881.1.S1_at, OsAffx.10020.1.S1_x_at, Os.48441.1.S1_at
	Os.51533.1.S1_at*, OsAffx.27824.1.S1_at, Os.53553.1.S1_at, OsAffx.15765.1.S1_at, Os.8956.1.S1_at
	Os.37062.2.S1_at*, OsAffx.29191.1.S1_at

Probe sets marked with * are used as the representative probe set for the cluster in the succeeding processes in the framework.

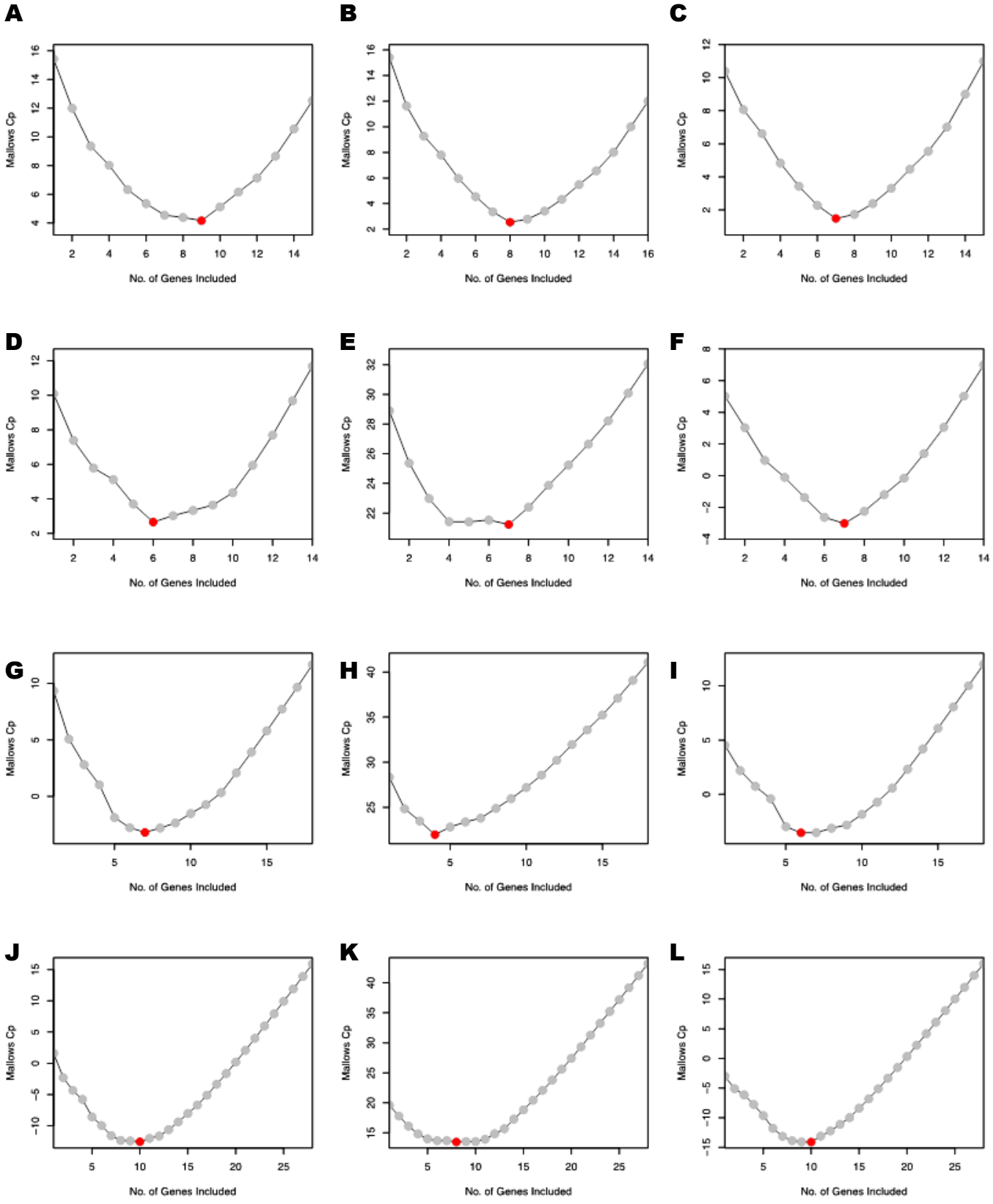

Figure 2. Plots of Mallows Cp values from the regression analysis. The y-axis represents the Mallows Cp values, while the x-axis represents the number of genes included in the regression model. The minimum of each plot is indicated by the red dot. **A**, **B**, and **C** represent Mallows Cp plots from regression analysis done on FL478 down-regulated genes where the linear model includes both QTLs, includes qSNK1 only, and includes qSNK9 only, respectively. **D**, **E**, and **F** are for FL478 up-regulated genes. **G**, **H**, and **I** are for IR29 down-regulated genes. **J**, **K**, and **L** are for IR29 up-regulated genes.

Table 6: FL478 down-regulated probe sets included in the linear model with varying pre-included QTLs.

Probe Set ID	Included in Model with:		
	qSNK1 Only	qSNK9 Only	qSNK1+qSNK9
OsAffx.13586.2.S1_at[a]	YES	YES	YES
Os.29814.1.S1_at[a]	YES	YES	YES
OsAffx.15286.1.S1_at[b]	YES	YES	YES
OsAffx.32296.1.A1_at[b]	YES	YES	YES
OsAffx.7216.1.S1_at[c]	YES	YES	YES
OsAffx.15760.1.S1_at[c]	YES	YES	YES
Os.6345.1.S1_at[c]	YES	YES	YES
Os.1503.1.S1_at[d]	YES	YES	YES
OsAffx.5111.1.S1_x_at[d]	YES	YES	YES
OsAffx.16888.1.S1_at[e]	YES	YES	YES
OsAffx.5782.1.S1_at[e]	YES	YES	YES
Os.10765.1.S1_at	YES	YES	YES
OsAffx.15397.1.S1_x_at[f]	YES	NO	NO
Os.53094.1.S1_at[f]	YES	NO	NO
Os.17112.1.S1_at[f]	YES	NO	NO
OsAffx.18135.1.S1_at	YES	NO	NO
OsAffx.19547.1.S1_at	YES	NO	NO
Os.34782.1.S1_at[g]	NO	YES	NO
OsAffx.21835.1.S1_x_at[g]	NO	YES	NO
OsAffx.30032.2.S1_at	NO	YES	NO
OsAffx.22071.1.S1_at	NO	NO	YES

Probe sets with the same letter superscript indicate that they belong to the same cluster formed during the first phase of multicollinearity handling.

Table 7: FL478 up-regulated probe sets included in the linear model with varying pre-included QTLs.

Probe Set ID	Included in Model with:		
	qSNK1 Only	qSNK9 Only	qSNK1+qSNK9
OsAffx.27871.1.S1_at[a]	YES	YES	YES
OsAffx.15672.1.S1_at[a]	YES	YES	YES
OsAffx.15353.1.S1_at[b]	YES	YES	YES
Os.4291.1.S1_at[b]	YES	YES	YES
OsAffx.15322.1.S1_x_at[b]	YES	YES	YES
OsAffx.30860.1.S1_at	YES	YES	YES
Os.39552.1.A1_s_at	YES	YES	YES
OsAffx.25794.1.S1_at	YES	NO	YES
OsAffx.18144.1.S1_at	YES	NO	NO
OsAffx.30227.1.S1_at	NO	YES	YES

continued on the next page

continuation of Table 7

Probe Set ID	Included in Model with:		
	qSNK1 Only	qSNK9 Only	qSNK1+qSNK9
OsAffx.30030.2.S1_at	NO	YES	YES
OsAffx.1477.1.A1_at	NO	YES	NO

Probe sets with the same letter superscript indicate that they belong to the same cluster formed during the first phase of multicollinearity handling.

Table 8: IR29 down-regulated probe sets included in the linear model with varying pre-included QTLs.

Probe Set ID	Included in Model with:		
	qSNK1 Only	qSNK9 Only	qSNK1+qSNK9
OsAffx.4155.1.S1_at	YES	YES	YES
OsAffx.24637.2.S1_x_at[a]	YES	NO	YES
OsAffx.8723.1.S1_at[a]	YES	NO	YES
OsAffx.15417.1.A1_at[a]	YES	NO	YES
OsAffx.4851.1.S1_x_at[a]	YES	NO	YES
OsAffx.4918.1.S1_at[a]	YES	NO	YES
OsAffx.15917.1.S1_at[b]	YES	NO	YES
OsAffx.5047.1.S1_at[b]	YES	NO	YES
OsAffx.4989.1.S1_at[c]	YES	NO	YES
OsAffx.5019.1.S1_at[c]	YES	NO	YES
OsAffx.25807.1.S1_at[d]	YES	NO	YES
Os.47942.1.A1_at[d]	YES	NO	YES
OsAffx.8455.1.S1_at	YES	NO	YES
OsAffx.5780.1.A1_at	YES	NO	NO
Os.23264.1.A1_at	NO	YES	NO
OsAffx.30902.1.S1_at	NO	YES	NO
Os.12750.1.S1_x_at	NO	YES	NO

Probe sets with the same letter superscript indicate that they belong to the same cluster formed during the first phase of multicollinearity handling.

Table 9: IR29 up-regulated probe sets included in the linear model with varying pre-included QTLs.

Probe Set ID	Included in Model with:		
	qSNK1 Only	qSNK9 Only	qSNK1+qSNK9
Os.52577.1.S1_x_at[a]	YES	YES	YES
Os.8413.2.A1_at[a]	YES	YES	YES
Os.8413.2.A1_x_at[a]	YES	YES	YES
Os.8413.2.A1_a_at[a]	YES	YES	YES
Os.11469.1.S1_at[a]	YES	YES	YES
Os.52577.1.S1_at[a]	YES	YES	YES
Os.26486.1.S1_at	YES	YES	YES

continued on the next page

continuation of Table 9

Probe Set ID	Included in Model with:		
	qSNK1 Only	qSNK9 Only	qSNK1+qSNK9
Os.55786.1.S1_at[b]	YES	YES	YES
Os.14214.1.S1_at[b]	YES	YES	YES
Os.31468.2.S1_at[b]	YES	YES	YES
Os.5265.1.S1_x_at[b]	YES	YES	YES
Os.4449.2.S1_at[b]	YES	YES	YES
Os.38806.1.A1_x_at[b]	YES	YES	YES
OsAffx.27513.1.S1_s_at[b]	YES	YES	YES
OsAffx.5120.1.S1_x_at[c]	YES	YES	YES
OsAffx.5111.1.S1_x_at[c]	YES	YES	YES
Os.27247.1.S1_at[c]	YES	YES	YES
Os.55116.1.S1_at[c]	YES	YES	YES
OsAffx.5122.1.S1_at[c]	YES	YES	YES
Os.27205.1.S1_at[c]	YES	YES	YES
Os.20817.5.A1_at	YES	YES	NO
OsAffx.450.1.S1_at	YES	YES	NO
Os.11330.1.S2_at	YES	NO	YES
Os.2957.1.S1_at	YES	NO	YES
Os.320.2.S1_a_at	YES	NO	YES
OsAffx.27821.1.S1_at[d]	YES	NO	YES
OsAffx.4896.1.S1_at[d]	YES	NO	YES
OsAffx.27747.1.S1_at[d]	YES	NO	YES
Os.27155.1.S1_at[d]	YES	NO	YES
OsAffx.15535.1.S1_at[d]	YES	NO	YES
Os.27790.1.A1_at[d]	YES	NO	YES
OsAffx.15364.1.S1_at[d]	YES	NO	YES
Os.4631.1.S1_at[d]	YES	NO	YES
Os.51533.1.S1_at[e]	YES	NO	YES
OsAffx.27824.1.S1_at[e]	YES	NO	YES
Os.53553.1.S1_at[e]	YES	NO	YES
OsAffx.15765.1.S1_at[e]	YES	NO	YES
Os.8956.1.S1_at[e]	YES	NO	YES
Os.49855.1.S1_at	NO	YES	NO
Os.24471.1.S1_at	NO	YES	NO
Os.17906.1.S1_at	NO	NO	YES
OsAffx.31987.1.S1_x_at	NO	NO	YES

Probe sets with the same letter superscript indicate that they belong to the same cluster formed during the first phase of multicollinearity handling.

DISCUSSION

A previous research (Walia et al. 2005) has already provided annotations as well as detailed discussion of the role of some of those genes included in the analysis. In addition to the annotations and descriptions provided by the literature, other information on the inferred causal genes was sought using the RiceChip site (http://ricechip.org) and the Rice Genome Annotation Project site (http://rice.plantbiology.msu.edu/). Tables 10-13 provide a summary of the annotations from the literature and from the aforementioned sites.

Varied selection of genes depending on QTLs included in the model

One interesting observation from the results is the selection of several genes as seemingly dependent on the QTLs included in the linear model. For instance, there were several genes that were chosen in multiple model scenarios based on the QTLs included during regression. Inclusion of those genes in multiple scenarios might connote that they affect salt tolerance invariant of the state of the QTLs. On the other hand, there were also genes that were included in specific regression model scenarios. This might suggest that certain genes react to a specific configuration of states of the QTLs. However, as mentioned in a previous section, the framework does not aim to establish causal dependence, or lack thereof, between QTLs and genes. Another test, or series of tests, *e.g.*, eQTL mapping and pleiotropy tests as done in Schadt et al. (2005), may be done in order to establish such causal relationship.

Characterization of inferred candidate causal genes

Characterization of the role of such genes to salt tolerance warrants a look into the annotations of those genes. For this particular study, the main basis for characterizing the inferred candidate causal genes would be the gene ontology (GO) terms included in the annotations.

The first set of genes to look at would be those which were identified to be related to stress response as per their GO terms. Genes related to stress response based on their GO terms are Os.39552.1.A1_s_at, OsAffx.25794.1.S1_at, Os.47942.1.A1_at, Os.23264.1.A1_at, Os.8413.2.A1_at, Os.8413.2.A1_x_at, Os.8413.2.A1_a_at, Os.5265.1.S1_x_at, Os.2957.1.S1_at, and Os.27155.S1_at. There were genes that were specifically annotated to be associated with response to abiotic stimulus, which includes salt stress. OsAffx.30030.2.S1_at, Os.47942.1.A1_at, Os.8413.2.A1_at, Os.8413.2.A1_x_at, Os.8413.2.A1_a_at, Os.27155.S1_at, and Os.15765.1.S1_at make up the list of those genes from the list of candidate causal genes that are related to abiotic stimulus response.

Genes pertinent to cell wall- and membrane-associated functions have also been associated with salt tolerance (Negrão et al. 2011, Walia et al. 2005, Zhang et al. 2012). In particular, cell wall- and membrane-related functions are important in the transport and regulation of Na^+ and K^+ within and among cells (Negrão et al. 2011). Candidate causal genes with cell wall- and membrane-associated functions are Os.29814.1.S1_at, OsAffx.32296.1.A1_at, Os15760.1.S1_at, Os.39552.1.A1_s_at, OsAffx.30030.2.S1_at, OsAffx.24637.2.S1_x_at, OsAffx.5047.1.S1_at, Os.27205.1.S1_at, Os.11469.1.S1_at, Os.26486.1.S1_at, Os.55786.1.S1_at, Os.5265.1.S1_x_at, Os.55116.1.S1_at, OsAffx.15364.1.S1_at, Os.8956.1.S1_at, and Os.17906.1.S1_at.

Signal transduction-related genes are also of importance to salt tolerance (Negrão et al. 2011, Zhang et al. 2012). In fact, studies suggest that response to abiotic stimulus in plants begins with signal transduction (Bouche et al. 2005, DeFalco et al. 2010, Negrão et al. 2011). OsAffx.5111.1.S1_x_at, OsAffx.5047.1.S1_at, and Os.5265.1.S1_x_at are among the candidate genes that are explicitly identified to be related to signal transduction. Transferases are also considered relevant components to signal transduction (Negrão et al. 2011, Zhang et al. 2012). A transferase is an enzyme that catalyzes transfer of functional groups between molecules. Genes identified to be related to transferase activity include OsAffx.32296.1.A1_at, Os.10765.1.S1_at, OsAffx.19547.1.S1_at, Os.52577.1.S1_x_at, Os.52577.1.S1_at, Os.55786.1.S1_at, OsAffx.27513.1.S1_s_at, Os.27205.1.S1_at, Os.27155.1.S1_at, OsAffx.15364.1.S1_at, and Os.24471.1.S1_at. A specific type of transferase is kinase, which facilitates activation of certain dormant substances. In the context of signal transduction, kinases are important in that they allow processes along a signaling pathway to proceed. Kinase activity-associated genes in the set of candidate causal genes are OsAffx.15760.1.S1_at, OsAffx.5111.1.S1_x_at, OsAffx.25794.1.S1_at, OsAffx.5047.1.S1_at, Os.23264.1.A1_at, and Os.49855.1.S1_at.

Hypothetical, unclassified, and putatively expressed proteins

Finally, there are those genes whose annotations only state that they putatively or hypothetically produce proteins. Walia et al. (2005) and even the RiceChip and Rice Genome Annotation Project search results did not provide additional insight as to the nature of such genes. Although no outright value can be drawn from these genes due to their lack of detailed annotations, their inclusion suggests that these genes might still have some role in salt tolerance, compounded by the fact that these genes were also found to be significantly differentially expressed, it can thus be suggested that further analysis as to the detailed nature of these genes be done.

CONCLUSIONS AND RECOMMENDATIONS

Similarly, if not more, essential to identifying genes related to a phenotypic trait is to identify those genes which cause the expression of the trait. The partial regression coefficient analysis framework proposed in this article provides the means for hy-

Table 10: Differential expression test p-values and annotations of FL478 down-regulated probe sets included in the regression analysis.

Probe Set	D.E. Test p-value	Walia et al. Annotation	Web-Sourced Annotation	PFAMs	Cellular Component GO	Biological Process GO	Molecular Function GO
OsAffx.13586.2.S1_at	0.039291	CDS expressed protein	Expressed protein	-	-	-	-
Os.29814.1.S1_at	0.042437	CDS putative beta-1,3-glucanase	glucan endo-1,3-beta-glucosidase precursor, putative, expressed	-	cellular_component (GO:0005575), cell wall (GO:0005618)	carbohydrate metabolic process (GO:0005975), metabolic process (GO:0008152)	catalytic activity (GO:0003824), hydrolase activity (GO:0016787), binding IEA (GO:0005488)
OsAffx.15286.1.S1_at	0.016845	CDS hypothetical protein	Hypothetical protein	-	-	-	-
OsAffx.32296.1.A1_at	0.001236	CDS 1,3-beta-D-glucan synthase, putative	1,3-beta-glucan synthase component domain containing protein, expressed	PF02364 PF04652	plasma membrane (GO:0005886), membrane (GO:0016020)	cellular process (GO:0009987), cell cycle (GO:0007049), development (GO:0007275), cell growth (GO:0016049), cell differentiation (GO:0030154), cell organization and biogenesis (GO:0016043), biosynthesis (GO:0009058), carbohydrate metabolism (GO:0005975), morphogenesis (GO:0009653), pollination (GO:0009856), metabolism (GO:0008152)	transferase activity (GO:0016740)
OsAffx.7216.1.S1_at	0.04065	CDS retrotransposon protein, putative, Ty3-gypsy sub-class	Retrotransposon, putative, centromere-specific	-	-	-	-
OsAffx.15760.1.S1_at	0.01161	CDS D-mannose binding lectin, putative	lectin protein kinase family protein, putative, expressed	-	plasma membrane (GO:0005886)	protein modification process (GO:0006464), cellular process (GO:0009987), metabolic process (GO:0008152)	protein binding (GO:0005515), kinase activity (GO:0016301), carbohydrate binding (GO:0030246)
Os.6345.1.S1_at	0.027485	CDS expressed protein	inositol oxygenase, putative, expressed	PF05153	cytoplasm (GO:0005737)	cellular process (GO:0009987), anatomical structure morphogenesis (GO:0009653), metabolic process (GO:0008152)	catalytic activity (GO:0003824)
Os.1503.1.S1_at	0.030353	CDS RISBZ5	BZIP transcription factor domain containing protein, expressed	PF00170	nucleus (GO:0005634)	nucleobase, nucleoside, nucleotide and nucleic acid metabolism (GO:0006139), biosynthesis (GO:0009058)	protein binding (GO:0005515), transcription factor activity (GO:0003700), DNA binding (GO:0003677)

continued on the next page

continuation of Table 10

Probe Set	D.E. Test p-value	Walia et al. Annotation	Web-Sourced Annotation	PFAMs	Cellular Component GO	Biological Process GO	Molecular Function GO
OsAffx.5111.1.S1_x_at	0.002262	CDS receptor-like protein kinase PRK1 - tomato	Inactive receptor kinase At2g26730 precursor, putative, expressed	PF00560 PF07714	-	protein modification (GO:0006464), signal transduction (GO:0007165)	kinase activity (GO:0016301), nucleotide binding (GO:0000166)
OsAffx.16888.1.S1_at*	0.025982	CDS hypothetical protein	Sym: ANAC001, NAC001,NAC domain containing protein 1, chr1:3760-5630 FORWARD LENGTH=1290	PF02365	cellular_component (GO:0005575)	development (GO:0007275), regulation of transcription, DNA-dependent (GO:0006355)	transcription factor activity (GO:0003700)
OsAffx.5782.1.S1_at	0.013959	CDS hypothetical protein	retrotransposon protein, putative, unclassified, expressed	-	-	-	-
Os.10765.1.S1_at	0.027956	CDS Transferase family	Transferase family protein putative, expressed	PF02458	cytosol (GO:0005829)	secondary metabolism (GO:0019748), biological_process (GO:0008150), cellular process (GO:0009987), biosynthesis (GO:0009058), metabolism (GO:0008152)	transferase activity (GO:0016740)
OsAffx.15397.1.S1_x_at*	0.03014	CDS hypothetical protein	Hydroxyproline-rich glycoprotein family protein, chr4:6952799-6954242 FORWARD LENGTH=660	-	cellular_component (GO:0005575)	biological_process (GO:0008150)	molecular_function (GO:0003674)
Os.53094.1.S1_at	0.035372	CDS hypothetical protein	expressed protein	-	cell (GO:0005623)	biological_process (GO:0008150)	-
Os.17112.1.S1_at	0.008202	CDS Similar to nicotianamine aminotransferase A	leucoanthocyanidin dioxygenase, putative, expressed	-	-	metabolic process (GO:0008152)	catalytic activity (GO:0003824), binding (GO:0005488)
OsAffx.18135.1.S1_at	0.005787	CDS hypothetical protein	-	-	-	-	-
OsAffx.19547.1.S1_at	0.004606	CDS Similar to dna-directed rna polymerase ii 8.2 kda polypeptide (ec 2.7.7.6) (rpb10) (rp10) (abc10). [mouse-ear cress	RNA polymerases N 8 kDa subunit, putative, expressed	PF01194	-	biosynthesis (GO:0009058), nucleobase, nucleoside, nucleotide and nucleic acid metabolism (GO:0006139)	binding (GO:0005488), transferase activity (GO:0016740), DNA binding (GO:0003677)

continued on the next page

continuation of Table 10

Probe Set	D.E. Test p-value	Walia et al. Annotation	Web-Sourced Annotation	PFAMs	Cellular Component GO	Biological Process GO	Molecular Function GO
Os.34782.1.S1_at	0.0106	CDS expressed protein	ARGOS, putative, expressed	-	-	-	-
OsAffx.21835.1.S1_x_at	0.021548	CDS Glutaredoxin, putative	OsGrx_C6 - glutaredoxin subgroup III, expressed	PF00462	-	cellular homeostasis (GO:0019725), metabolic process (GO:0008152)	molecular_function (GO:0003674), catalytic activity (GO:0003824)
OsAffx.30032.2.S1_at	0.00177	CDS hypothetical protein	-	-	-	-	-
OsAffx.22071.1.S1_at	0.024348	CDS Protein kinase domain, putative	-	-	-	-	-

The differential expression (D.E.) test p-values in the second column and the Walia et al. annotations were taken from the Walia et al. (2005) article. Web-sourced annotations were those retrieved from either the RiceChip or the Rice Genome Annotation Project site. Probe set IDs marked with * have web-sourced annotations, PFAMs, and GO terms that were from best tBLASTx match of *Oryza* model to *Arabidopsis*, as no existing annotation was retrieved for rice. Cells with - as values have no retrieved entries for the respective column.

Table 11: Differential expression test p-values and annotations of FL478 up-regulated probe sets included in the regression analysis.

Probe Set	D.E. Test p-value	Walia et al. Annotation	Web-Sourced Annotation	PFAMs	Cellular Component GO	Biological Process GO	Molecular Function GO
OsAffx.27871.1.S1_at	0.001523	CDS Putative far-red impaired response protein	Far1-like, putative	-	-	-	-
OsAffx.15672.1.S1_at*	0.0025	CDS MATE efflux family protein, putative	TTF-type zinc finger protein with HAT dimerisation domain, chr1:6657260-6659569 REVERSE LENGTH=2310	PF05699	plasmodesma (GO:0009506)	biological_process (GO:0008150)	protein dimerization activity (GO:0046983)
OsAffx.15353.1.S1_at*	0.043997	CDS hypothetical protein	Sym: ANAC001, NAC001, NAC domain containing protein 1, chr1:3760-5630 FORWARD LENGTH=1	PF02365	cellular_component (GO:0005575)	development (GO:0007275), regulation of transcription, DNA-dependent (GO:0006355)	transcription factor activity (GO:0003700)
Os.4291.1.S1_at	0.025458	CDS Cytochrome P450	cytochrome P450 86A1, putative, expressed	-	endoplasmic reticulum (GO:0005783)	metabolic process (GO:0008152)	oxygen binding (GO:0019825), catalytic activity (GO:0003824), binding (GO:0005488), molecular_function (GO:0003674)

continued on the next page

continuation of Table 11

Probe Set	D.E. Test p-value	Walia et al. Annotation	Web-Sourced Annotation	PFAMs	Cellular Component GO	Biological Process GO	Molecular Function GO
OsAffx.15322.1.S1_x_at	0.030984	CDS Similar to cyto-chrome P450	expressed protein	-	cell (GO:0005623)	metabolic process (GO:0008152)	oxygen binding (GO:0019825), catalytic activity (GO:0003824), binding (GO:0005488), molecular_function (GO:0003674)
OsAffx.30860.1.S1_at	0.001294	CDS hypothetical protein	expressed protein	-	-	-	-
Os.39552.1.A1_s_at	0.042304	CDS protein phosphatase 2C	Protein phosphatase 2C, putative, expressed	PF00481	plasma membrane (GO:0005886)	biological_process (GO:0008150), response to stress (GO:0006950), response to biotic stimulus (GO:0009607), protein modification (GO:0006464), cellular process (GO:0009987), metabolism (GO:0008152)	hydrolase activity (GO:0016787)
OsAffx.25794.1.S1_at*	0.03654	CDS putative protein kinase	Protein kinase protein with adenine nucleotide alpha hydrolases-like domain, chr5:25588254-25591229 FORWARD LENGTH=2118	PF00069 PF00582	-	protein amino acid phosphorylation (GO:0006468), response to stress (GO:0006950), phosphorylation (GO:0016310)	kinase activity (GO:0016301), protein serine/threonine kinase activity (GO:0004674), protein kinase activity (GO:0004672), ATP binding (GO:0005524)
OsAffx.18144.1.S1_at	0.004246	CDS hypothetical protein	Expressed protein	-	-	-	-
OsAffx.30227.1.S1_at	0.00269	CDS hypothetical protein	Expressed protein	-	-	-	-
OsAffx.30030.2.S1_at	0.002768	CDS Avr9/Cf-9 rapidly elicited protein 141	Glutamate receptor, putative, expressed	PF00060 PF00497 PF01094	cell (GO:0005623), membrane (GO:0016020)	cell homeostasis (GO:0019725), cellular process (GO:0009987), response to abiotic stimulus (GO:0009628), transport (GO:0006810)	transporter activity (GO:0005215)
OsAffx.1477.1.A1_at	0.040795	CDS retrotransposon protein, putative, Ty3-gypsy sub-class	-	-	-	-	-

The differential expression (D.E.) test p-values in the second column and the Walia et al. annotations were taken from the Walia et al. (2005) article. Web-sourced annotations were those retrieved from either the RiceChip or the Rice Genome Annotation Project site. Probe set IDs marked with * have web-sourced annotations, PFAMs, and GO terms that were from best tBLASTx match of Oryza model to Arabidopsis, as no existing annotation was retrieved for rice. Cells with - as values have no retrieved entries for the respective column.

Table 12: Differential expression test p-values and annotations of IR29 down-regulated probe sets included in the regression analysis.

Probe Set	D.E. Test p-value	Walia et al. Annotation	Web-Sourced Annotation	PFAMs	Cellular Component GO	Biological Process GO	Molecular Function GO
OsAffx.4155.1.S1_at	0.037186	CDS hypothetical protein	Expressed protein	-	-	-	-
OsAffx.24637.2.S1_x_at*	0.000531	CDS hypothetical protein	Leucine-rich repeat (LRR) family protein, chr4:7758610-7760892 FORWARD LENGTH=2283	PF08263	cell wall (GO:0005618), cell wall (sensu Magnoliophyta) (GO:0009505)	-	protein binding (GO:0005515)
OsAffx.8723.1.S1_at	0.005944	CDS hypothetical protein	-	-	-	-	-
OsAffx.15417.1.A1_at	0.014466	CDS hypothetical protein	-	-	-	-	-
OsAffx.4851.1.S1_x_at	0.028435	CDS hypothetical protein	expressed protein	-	-	-	-
OsAffx.4918.1.S1_at	0.046027	CDS hypothetical protein	expressed protein	-	-	-	-
OsAffx.15917.1.S1_at	0.003645	CDS hypothetical protein	Expressed protein	-	-	-	-
OsAffx.5047.1.S1_at	0.021818	CDS Leucine Rich Repeat, putative	Receptor-like protein kinase precursor, putative, expressed	PF00069 PF00560 PF08263	plasma membrane (GO:0005886)	signal transduction (GO:0007165), protein modification (GO:0006464)	kinase activity (GO:0016301), nucleotide binding (GO:0000166)
OsAffx.4989.1.S1_at	0.005598	CDS hypothetical protein	Expressed protein	-	-	-	-
OsAffx.5019.1.S1_at	0.01806	CDS AT4g20860/ T13K14_20	reticuline oxidase-like protein precursor, putative, expressed	-	cytosol (GO:0005829)	metabolic process (GO:0008152)	catalytic activity (GO:0003824), binding (GO:0005488), molecular_function (GO:0003674)
OsAffx.25807.1.S1_at	0.002187	CDS hypothetical protein	expressed protein	-	-	-	-
Os.47942.1.A1_at	0.04037	CDS putative calreticulin precursor	Calreticulin precursor, putative, expressed	PF00262	mitochondrion (GO:0005739), vacuole (GO:0005773), endoplasmic reticulum (GO:0005783)	response to abiotic stimulus (GO:0009628), response to stress (GO:0006950)	binding (GO:0005488), protein binding (GO:0005515)
OsAffx.8455.1.S1_at	0.030958	CDS transposon protein, putative, CACTA, En/ Spm sub-class	transposon protein, putative, CACTA, En/Spm sub-class, expressed	-	-	-	-
OsAffx.5780.1.A1_at	0.016749	CDS hypothetical protein	-	-	-	-	-
Os.23264.1.A1_at	0.007744	CDS Protein kinase domain, putative	TKL_IRAK_DUF26-la.5 - DUF26 kinases have homology to DUF26 containing loci, expressed	PF01657 PF07714	cell (GO:0005623)	protein modification (GO:0006464), response to stress (GO:0006950)	nucleotide binding (GO:0000166), kinase activity (GO:0016301)
OsAffx.30902.1.S1_at	0.036331	CDS F-box domain, putative	-	-	-	-	-

continued on the next page

continuation of Table 12

Probe Set	D.E. Test p-value	Walia et al. Annotation	Web-Sourced Annotation	PFAMs	Cellular Component GO	Biological Process GO	Molecular Function GO
Os.12750.1.S1_x_at	0.037186	CDS K-box region, putative	OsMADS15 - MADS-box family gene with MIKCc type-box, expressed	PF01486 PF00319	nucleus (GO:0005634)	anatomical structure morphogenesis (GO:0009653), multicellular organismal development (GO:0007275), cell differentiation (GO:0030154), flower development (GO:0009908), nucleobase, nucleoside, nucleotide and nucleic acid metabolic process (GO:0006139), biosynthetic process (GO:0009058)	sequence-specific DNA binding transcription factor activity (GO:0003700), protein binding (GO:0005515), DNA binding (GO:0003677)

The differential expression (D.E.) test p-values in the second column and the Walia et al. annotations were taken from the Walia et al. (2005) article. Web-sourced annotations were those retrieved from either the RiceChip or the Rice Genome Annotation Project site. Probe set IDs marked with * have web-sourced annotations, PFAMs, and GO terms that were from best tBLASTx match of *Oryza* model to *Arabidopsis*, as no existing annotation was retrieved for rice. Cells with - as values have no retrieved entries for the respective column.

Table 13: Differential expression test p-values and annotations of IR29 up-regulated probe sets included in the regression analysis.

Probe Set	D.E. Test p-value	Walia et al. Annotation	Web-Sourced Annotation	PFAMs	Cellular Component GO	Biological Process GO	Molecular Function GO
Os.52577.1.S1_x_at	0.001972	CDS UDP-glucoronosyl and UDP-glucosyl transferase	Anthocyanin 3-O-beta-glucosyltransferase, putative, expressed	PF00201	cell (GO:0005623)	metabolism (GO:0008152)	transferase activity (GO:0016740)
Os.8413.2.A1_at	0.00366	CDS expressed protein	Male sterility protein, putative, expressed	PF03015P F07993	plastid (GO:0009536)	development (GO:0007275), response to stress (GO:0006950), response to abiotic stimulus (GO:0009628), metabolism (GO:0008152), biosynthesis (GO:0009058), cellular process (GO:0009987), lipid metabolism (GO:0006629), secondary metabolism (GO:0019748), cell cycle (GO:0007049)	catalytic activity (GO:0003824)

continued on the next page

continuation of Table 13

Probe Set	D.E. Test p-value	Walia et al. Annotation	Web-Sourced Annotation	PFAMs	Cellular Component GO	Biological Process GO	Molecular Function GO
Os.8413.2.A1_x_at	0.009643	CDS expressed protein	male sterility protein, putative, expressed	-	plastid (GO:0009536)	response to abiotic stimulus (GO:0009628), cell cycle (GO:0007049), multicellular organismal development (GO:0007275), secondary metabolic process (GO:0019748), lipid metabolic process (GO:0006629), cellular process (GO:0009987), biosynthetic process (GO:0009058), metabolic process (GO:0008152), response to stress (GO:0006950)	catalytic activity (GO:0003824)
Os.8413.2.A1_a_at	0.005706	CDS expressed protein	male sterility protein, putative, expressed	-	plastid (GO:0009536)	response to abiotic stimulus (GO:0009628), cell cycle (GO:0007049), multicellular organismal development (GO:0007275), secondary metabolic process (GO:0019748), lipid metabolic process (GO:0006629), cellular process (GO:0009987), biosynthetic process (GO:0009058), metabolic process (GO:0008152), response to stress (GO:0006950)	catalytic activity (GO:0003824)
Os.11469.1.S1_at	0.006539	CDS oxidoreductase, aldo/keto reductase family	oxidoreductase, aldo/keto reductase family protein, putative, expressed	PF00248	plasma membrane (GO:0005886), plastid (GO:0009536)	metabolic process (GO:0008152)	catalytic activity (GO:0003824)
Os.52577.1.S1_at	0.007597	CDS UDP-glucoronosyl and UDP-glucosyl transferase	anthocyanin 3-O-beta-glucosyltransferase, putative, expressed	-	cell (GO:0005623)	metabolic process (GO:0008152)	transferase activity (GO:0016740)
Os.26486.1.S1_at	0.016088	CDS Gl1 protein	WAX2, putative, expressed	PF04116, PF12076	plasma membrane (GO:0005886), membrane (GO:0016020)	metabolism (GO:0008152), development (GO:0007275), cell differentiation (GO:0030154), biosynthesis (GO:0009058), lipid metabolism (GO:0006629), cellular process (GO:0009987), reproduction (GO:0000003)	catalytic activity (GO:0003824), binding (GO:0005488)
Os.55786.1.S1_at	0.016223	CDS hypothetical protein	Hypro1, putative, expressed	PF01762	cell (GO:0005623), membrane (GO:0016020)	biosynthesis (GO:0009058), protein modification (GO:0006464), carbohydrate metabolism (GO:0005975), metabolism (GO:0008152)	transferase activity (GO:0016740)

continued on the next page

continuation of Table 13

Probe Set	D.E. Test p-value	Walia et al. Annotation	Web-Sourced Annotation	PFAMs	Cellular Component GO	Biological Process GO	Molecular Function GO
Os.14214.1.S1_at	0.020276	CDS Eukaryotic aspartyl protease, putative	aspartic proteinase nepenthesin-2 precursor, putative, expressed	-	-	protein metabolic process (GO:0019538)	hydrolase activity (GO:0016787), DNA binding (GO:0003677)
Os.31468.2.S1_at	0.035048	CDS hypothetical protein	expressed protein	-	cell (GO:0005623)	protein modification process (GO:0006464), cellular process (GO:0009987), catabolic process (GO:0009056), protein metabolic process (GO:0019538)	catalytic activity (GO:0003824), protein binding (GO:0005515)
Os.5265.1.S1_x_at	0.033296	CDS expressed protein	flavin-containing monooxygenase family protein, putative, expressed	-	endoplasmic reticulum (GO:0005783), membrane (GO:0016020), cell (GO:0005623)	cell death (GO:0008219), signal transduction (GO:0007165), metabolic process (GO:0008152), response to biotic stimulus (GO:0009607), cellular process (GO:0009987), response to stress (GO:0006950)	nucleotide binding (GO:0000166), catalytic activity (GO:0003824), binding (GO:0005488)
Os.4449.2.S1_at	0.047191	CDS hypothetical protein	-	-	-	-	-
Os.38806.1.A1_x_at	0.029355	CDS hypothetical protein	expressed protein	-	-	-	-
OsAffx.27513.1.S1_s_at	0.034509	CDS N-hydroxycinnamoyl/ benzoyltransferase	transferase family protein, putative, expressed	-	cytosol (GO:0005829)	biological_process (GO:0008150), secondary metabolic process (GO:0019748), cellular process (GO:0009987), biosynthetic process (GO:0009058), metabolic process (GO:0008152)	transferase activity (GO:0016740)
OsAffx.5120.1.S1_x_at	0.00911	CDS Cytochrome P450	Cytochrome P450, putative, expressed	PF00067	-	metabolism (GO:0008152)	binding (GO:0005488), catalytic activity (GO:0003824), molecular_function (GO:0003674), oxygen binding (GO:0019825)
OsAffx.5111.1.S1_x_at	0.015673	CDS receptor-like protein kinase PRK1 - tomato	inactive receptor kinase At2g26730 precursor, putative, expressed	-	-	protein modification process (GO:0006464), signal transduction (GO:0007165)	nucleotide binding (GO:0000166), kinase activity (GO:0016301)
Os.27247.1.S1_at	0.011867	CDS GDSL-like Lipase/ Acylhydrolase	GDSL-like lipase/ acylhydrolase, putative, expressed	-	cell (GO:0005623)	lipid metabolic process (GO:0006629), metabolic process (GO:0008152)	hydrolase activity (GO:0016787)

continued on the next page

continuation of Table 13

Probe Set	D.E. Test p-value	Walia et al. Annotation	Web-Sourced Annotation	PFAMs	Cellular Component GO	Biological Process GO	Molecular Function GO
Os.55116.1.S1_at	0.013683	CDS hypothetical protein	expressed protein	-	plasma membrane (GO:0005886)	biological_process (GO:0008150)	molecular_function (GO:0003674)
OsAffx.5122.1.S1_at	0.03333	CDS hypothetical protein	hypothetical protein	-	-	-	-
Os.27205.1.S1_at	0.043581	CDS xyloglucan endotransglycosylase homolog.	glycosyl hydrolases family 16, putative, expressed	-	cell wall (GO:0005618)	cellular process (GO:0009987), carbohydrate metabolic process (GO:0005975), metabolic process (GO:0008152)	hydrolase activity (GO:0016787), transferase activity (GO:0016740)
Os.20817.5.A1_at	0.011517	CDS Protein kinase domain, putative	Expressed protein	-	-	-	-
OsAffx.450.1.S1_at	0.041253	CDS retrotransposon protein, putative, Ty1-copia sub-class	retrotransposon protein, putative, Ty1-copia sub-class	-	-	-	-
Os.11330.1.S2_at*	0.009355	CDS hypothetical protein	Sym: ATDI21, DI21, drought-induced 21, chr4:9028657-9029269 FORWARD LENGTH=315	PF03242	-	response to water deprivation (GO:0009414), embryonic development (GO:0009790), response to abscisic acid stimulus (GO:0009737)	molecular_function (GO:0003674)
Os.2957.1.S1_at	0.021391	CDS putative peroxidase	peroxidase precursor, putative, expressed	-	cell (GO:0005623)	metabolic process (GO:0008152), response to stress (GO:0006950)	catalytic activity (GO:0003824), binding (GO:0005488)
Os.320.2.S1_a_at	0.028234	CDS hypothetical protein	OsMADS3 - MADS-box family gene with MIKCc type-box, expressed	-	cell (GO:0005623), nucleus (GO:0005634)	multicellular organismal development (GO:0007275), nucleobase, nucleoside, nucleotide and nucleic acid metabolic process (GO:0006139), biosynthetic process (GO:0009058), cell differentiation (GO:0030154), flower development (GO:0009908)	sequence-specific DNA binding transcription factor activity (GO:0003700), protein binding (GO:0005515), DNA binding (GO:0003677)
OsAffx.27821.1.S1_at	0.003885	CDS hypothetical protein	hypothetical protein	-	-	-	-
OsAffx.4896.1.S1_at	0.019731	CDS hypothetical protein	-	-	-	-	-
OsAffx.27747.1.S1_at	0.019941	CDS glycine rich protein	glycine rich protein family protein, putative, expressed	-	-	-	-

continued on the next page

continuation of Table 13

Probe Set	D.E. Test p-value	Walia et al. Annotation	Web-Sourced Annotation	PFAMs	Cellular Component GO	Biological Process GO	Molecular Function GO
Os.27155.1.S1_at	0.037566	CDS Chalcone and stilbene synthases, C-terminal domain, putative	3-ketoacyl-CoA synthase, putative, expressed	PF08541 PF08392	endoplasmic reticulum (GO:0005783)	biosynthetic process (GO:0009058), response to abiotic stimulus (GO:0009628), response to stress (GO:0006950), metabolic process (GO:0008152), multicellular organismal development (GO:0007275), lipid metabolic process (GO:0006629), cellular process (GO:0009987), cell growth (GO:0016049), anatomical structure morphogenesis (GO:0009653), cellular component organization (GO:0016043)	transferase activity (GO:0016740), catalytic activity (GO:0003824)
OsAffx.15535.1.S1_at	0.03339	CDS probable glycine-rich protein [imported] - Arabidopsis thaliana	glycine-rich cell wall structural protein precursor, putative, expressed	-	-	-	-
Os.27790.1.A1_at	0.040138	CDS hypothetical protein	expressed protein	-	-	-	-
OsAffx.15364.1.S1_at	0.04793	CDS Xyloglucan fucosyltransferase, putative	xyloglucan fucosyltransferase, putative, expressed	-	membrane (GO:0016020), cell (GO:0005623)	cellular process (GO:0009987), metabolic process (GO:0008152)	transferase activity (GO:0016740)
Os.4631.1.S1_at	0.047732	CDS expressed protein	expressed protein	-	-	-	-
Os.51533.1.S1_at	0.01016	CDS Zinc finger, C3HC4 type (RING finger), putative	zinc finger, C3HC4 type domain containing protein, expressed	-	-	-	binding (GO:0005488)
OsAffx.27824.1.S1_at	0.017432	CDS hypothetical protein	expressed protein	-	-	-	-
Os.53553.1.S1_at	0.048348	CDS Uncharacterized ACR, COG2135, putative	expressed protein	-	-	biological_process (GO:0008150)	molecular_function (GO:0003674)
OsAffx.15765.1.S1_at	0.031379	CDS flowering locus T	osFTL12 FT-Like12 homologous to Flowering Locus T gene; contains Pfam profile PF01161: Phosphatidylethanolamine-binding protein, expressed	-	cytoplasm (GO:0005737), nucleus (GO:0005634)	cellular process (GO:0009987), response to abiotic stimulus (GO:0009628), post-embryonic development (GO:0009791), reproduction (GO:0000003), flower development (GO:0009908)	lipid binding (GO:0008289), protein binding (GO:0005515)
Os.8956.1.S1_at	0.043411	CDS probable amino acid carrier [imported] - Arabidopsis thaliana	amino acid transporter, putative, expressed	-	membrane (GO:0016020)	cellular process (GO:0009987)	cellular process (GO:0009987), transport (GO:0006810)

continued on the next page

continuation of Table 13

Probe Set	D.E. Test p-value	Walia et al. Annotation	Web-Sourced Annotation	PFAMs	Cellular Component GO	Biological Process GO	Molecular Function GO
Os.49855.1.S1_at	0.003309	CDS Protein kinase domain, putative	Tyrosine protein kinase domain containing protein, putative, expressed	PF00069	-	development (GO:0007275), protein modification (GO:0006464), metabolism (GO:0008152), cellular process (GO:0009987)	kinase activity (GO:0016301), nucleotide binding (GO:0000166)
Os.24471.1.S1_at	0.029594	CDS UDP-glucoronosyl and UDP-glucosyl transferase	UDP-glucoronosyl and UDP -glucosyl transferase domain containing protein, expressed	-	-	metabolic process (GO:0008152)	transferase activity (GO:0016740)
Os.17906.1.S1_at	0.000697	CDS hypothetical protein	Isoflavone reductase, putative, expressed	PF05368	thylakoid (GO:0009579), plastid (GO:0009536), membrane (GO:0016020)	metabolism (GO:0008152)	catalytic activity (GO:0003824), binding (GO:0005488)
OsAffx.31987.1.S1_x_at	0.025405	CDS hypothetical protein	-	-	-	-	-

The differential expression (D.E.) test p-values in the second column and the Walia et al. annotations were taken from the Walia et al. (2005) article. Web-sourced annotations were those retrieved from either the RiceChip or the Rice Genome Annotation Project site. Probe set IDs marked with * have web-sourced annotations, PFAMs, and GO terms that were from best tBLASTx match of *Oryza* model to *Arabidopsis*, as no existing annotation was retrieved for rice. Cells with - as values have no retrieved entries for the respective column.

pothesis generation in that regard and can thus be a good initial step towards identifying genes causal to a phenotypic trait. The technique infers a set of candidate causal genes by assessing the inferred allelic state of the sequence of the genes, together with known QTLs, in relation to the phenotypic trait values, almost similar to what is done in QTL mapping techniques.

The use of the technique to infer candidate causal genes to rice salt tolerance had interesting results at the very least. Several genes related to physiological (cell wall- and membrane-related) and biochemical (signal transduction) mechanisms pertinent to salt tolerance have been inferred as potentially causal to the trait. Characterization of certain candidate causal genes suggests that the method can supply relevant prospective causal genes that are, at least, indeed related to salt tolerance.

One possible extension of the proposed framework entails allowing the methods to handle non-RIL cases, especially data sets whose marker score data contain significant numbers of heterozygous alleles. However, we must keep in mind that an advantage of using RILs is that the allelic states of the samples are mostly, if not all, homozygous, something which might not be the case in non-RILs. Another possible extension is to be able to infer potential causal relationships, or lack thereof, between QTLs and genes, as with Schadt et al. (2005), thus further strengthening the causality of gene expression to trait. This is considered a significant addendum taking into account the possibility that eQTL mapping may not be altogether possible, or the results may not be as definitive, due to the small number of gene expression data samples.

ACKNOWLEDGEMENTS

The corresponding author would like to thank the Department of Science and Technology's Engineering Research and Development for Technology (ERDT) program for funding this research. Also, many thanks to Dr. Michael Thomson, Ms. Marjorie de Ocampo, and Dr. Glenn Gregorio of IRRI for providing assistance in understanding the rice salt stress marker, QTL, and phenotypic data sets. Finally, much gratitude and appreciation is extended to the reviewers of the paper whose extensive critique and comments helped significantly improve the manuscript, and even part of the proposed method itself.

CONFLICTS OF INTEREST

Dr. Eduardo Mendoza, co-author of this paper, is part of the editorial board of the Philippine Science Letters.

CONTRIBUTIONS OF INDIVIDUAL AUTHORS

JM Yap developed the framework and the techniques detailed in the paper as well as performed the experiments and processed the results. He is also the main writer of the manuscript. R

Mauleon provided resources and "technical support" on rice data. He also provided technical details on the bioinformatics parts of the paper. E Mendoza provided technical details on the mathematics part of the paper and was the proofreader of the manuscript. H Adorna provided additional technical details on the mathematical and computational aspects of the paper.

REFERENCES:

Al-Shahrour F, Diaz-Uriarte R, Dopazo J. Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information. Bioinformatics 2005; 21(13): 2988-2993.

Bouche N, Yellin A, Snedden W, Fromm H. Plant-specific calmodulin-binding proteins. Annu Rev Plant Biol 2005; 56: 435-466.

Broman KW. The genomes of recombinant inbred lines. Genetics 2005; 169: 1133-1146.

Collard BCY, Jahufer MZZ, Brouwer JB, Pang ECK. An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. Euphytica 2005; 142(1-2): 169-196.

Chu J, Weiss ST, Carey VJ, Raby BA. A graphical model approach for inferring large-scale networks integrating gene expression and genetic polymorphism. BMC Syst Biol 2009; 3:55.

DeFalco T, Bender K, Snedden W. Breaking the code: Ca^{2+} sensors in plant signalling. Biochem J 2010; 425: 27-40.

Degnan JH, Lasky-Su J, Raby BA, Xu M, Molony C, Schadt EE, Lange C. Genomics and genome-wide association studies: An integrative approach for expression QTL mapping. Genomics 2008; 92(3): 129-133.

Druka A, Druka I, Centeno AG, Li H, Sun Z, Thomas WTB, Bonar N, Steffenson BJ, Ullrich SE, Kleinhofs A, Wise, RP, Close TJ, Potokina E, Luo Z, Wagner C, Schweizer GF, Marshall DF, Kearsey MJ, Williams RW, Waugh R. Towards system genetic analyses in barley: Integration of phenotypic, expression, and genotype data into GeneNetwork. BMC Genet 2008; 9: 73.

Kang HP, Yang X, Chen R, Zhang B, Corona E, Schadt EE, Butte AJ. Integration of disease-specific single nucleotide polymorphisms, expression quantitative trait loci and co-expression networks reveal novel candidate genes for type 2 diabetes. Diabetologia 2012; 55: 2205-2213.

Lumley T. leaps: regression subset selection. http://CRAN.R-project.org/package=leaps 2009.

Lusis AJ, Attie AD, Reue K. Metabolic syndrome: from epidemiology to systems biology. Nat Rev Genet 2008; 9: 819-830.

Miller AJ. Finding subsets which fit well. In: Isham V, Keiding N, Louis T, Reid N, Tibshirani R, Tong H, eds. Subset Selection in Regression. Second Edition. Boca Raton:Chapman and Hall/CRC, 2002: 37-85.

Negrão S, Courtois B, Ahmadi N, Abreu I, Saibo N, Oliviera MM. Recent updates on salinity stress in rice: from physiological to molecular responses. Crit Rev Plant Sci 2011; 30: 329-377.

Schadt E, Lamb J, Yang X, Zhu J, Edwards S, GuhaThakurta D, Sieberts SK, Monks S, Reitman M, Zhang C, Lum PY, Leonardson A, Thieringer R, Metzger JM, Yang L, Castle J, Zhu H, Kash SF, Drake TA, Sachs A, Lusis AJ. An integrative genomics approach to infer causal associations between gene expression and disease. Nat Genet 2005; 37 (7): 710-717.

Suits DB. Use of dummy variables in regression equations. J Am Statist Assoc 1957; 52(280): 548-551.

Thomson MJ, de Ocampo M, Egdane J, Rahman MA, Sajise AG, Adorada DL, Tumimbang-Raiz E, Blumwald E, Seraj ZI, Singh RK, Gregorio GB, Ismail AM. Characterizing the saltol quantitative trait locus for salinity tolerance in rice. Rice 2010; 3: 148-160.

Walia H, Wilson C, Condamine P, Liu X, Ismail AM, Zeng L, Wanamaker SI, Mandal J, Xu J, Cui X, Close TJ. Comparative transcriptional profiling of two contrasting rice genotypes under salinity stress during the vegetative growth stage. Plant Physiol 2005; 139(2): 822-835.

Zhang H, Han B, Wang T, Chen S, Li H, Zhang Y, Dai S. Mechanisms of plant salt response: insights from proteomics. J Proteome Res 2012; 11: 49-67.