Analytica Chimica Acta 705 (2011) 56–63

Contents lists available at ScienceDirect

Analytica Chimica Acta

journal homepage: www.elsevier.com/locate/aca

Data integration and network reconstruction with ~omics data using Random Forest regression in potato

Animesh Acharjee[a,b,*], Bjorn Kloosterman[b], Ric C.H. de Vos[c,d], Jeroen S. Werij[b,c], Christian W.B. Bachem[b], Richard G.F. Visser[b,c], Chris Maliepaard[b]

[a] Graduate School Experimental Plant Sciences, Wageningen, The Netherlands
[b] Wageningen UR Plant Breeding, Wageningen University and Research Center, P.O. Box 386, 6700 AJ Wageningen, The Netherlands
[c] Centre for BioSystems Genomics, P.O. Box 98, 6700 AA, Wageningen, The Netherlands
[d] Plant Research International, P.O. Box 16, 6700 AA Wageningen, The Netherlands

ARTICLE INFO

Article history:
Received 30 December 2010
Received in revised form 23 March 2011
Accepted 25 March 2011
Available online 13 April 2011

Keywords:
Data integration
Random Forest
Network reconstruction
Tuber flesh color
Potato

ABSTRACT

In the post-genomic era, high-throughput technologies have led to data collection in fields like transcriptomics, metabolomics and proteomics and, as a result, large amounts of data have become available. However, the integration of these ~omics data sets in relation to phenotypic traits is still problematic in order to advance crop breeding. We have obtained population-wide gene expression and metabolite (LC–MS) data from tubers of a diploid potato population and present a novel approach to study the various ~omics datasets to allow the construction of networks integrating gene expression, metabolites and phenotypic traits. We used Random Forest regression to select subsets of the metabolites and transcripts which show association with potato tuber flesh color and enzymatic discoloration. Network reconstruction has led to the integration of known and uncharacterized metabolites with genes associated with the carotenoid biosynthesis pathway. We show that this approach enables the construction of meaningful networks with regard to known and unknown components and metabolite pathways.

Crown Copyright © 2011 Published by Elsevier B.V. All rights reserved.

1. Introduction

High-throughput ~omics technologies like microarrays [1,2] mass spectrometry (*e.g.* LC–MS, GC–MS) [3,4] and protein chips [5,6] have gained much interest in the biological domain [26]. These techniques allow one to measure thousands of variables (genes, metabolites, proteins) simultaneously across populations. The data generated by these techniques: transcriptomics, metabolomics and proteomics, are often collectively denoted as ~omics data [7]. To understand the organization of cellular functions at different levels (gene, metabolite, or protein) and link them to a particular phenotype, an integrative approach is needed and is often referred to as "systems biology" [11,12]. Major challenges include interpretation and integration of large datasets to understand the principles underlying the regulation of genes, metabolites and proteins and also how their combined interactions associate with variation in phenotype [13,14].

Several attempts have been made to integrate multiple ~omics data sets from different species such as metabolomics and proteomics [22], and transcriptomics and metabolomics in *Arabidopsis thaliana* [21], transcriptomics and proteomics in soybean [24] and transcriptomics, metabolomics and proteomics in grapevine berry [23].

In this study, we integrate population wide transcriptomics and metabolomics data sets with observed variation in potato tuber quality traits to search for novel associations. Secondly, the genetic basis for these traits and their ~omics-data associations will be analyzed and we construct a correlation network of genes and metabolites associated with the quality traits.

Potato tubers are considered as an important and healthy addition to the human diet and therefore much effort has been undertaken to improve the accumulation of healthy compounds such as carotenoids. Carotenoids are thought to be the primary determinant of tuber flesh color [18] and in recent years, much progress has been made in the identification of key regulatory genes in potato tuber carotenoid content [19,30]. A second tuber quality trait important for consumers and processing industry is enzymatic discoloration. Quantitative trait loci (QTLs) [37] for flesh color and enzymatic discoloration and other quality traits have been reported [18,19,28]. The QTL analysis of both flesh color and enzymatic discoloration within a well studied diploid

* Corresponding author at: Wageningen UR Plant Breeding, Wageningen University and Research Center, P.O. Box 386, 6700 AJ Wageningen, The Netherlands. Tel.: +31 317 482815; fax: +31 317 483457.
E-mail address: animesh.acharjee@wur.nl (A. Acharjee).

0003-2670/$ – see front matter. Crown Copyright © 2011 Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.aca.2011.03.050

potato population (here denoted as C × E) shows both unique and overlapping QTLs across the genome. Most interesting is the strong overlap between a QTL for flesh color and enzymatic discoloration of tubers on chromosome 3.

In this study, we used Random Forest (RF) regression for integrating the transcriptomics and metabolomics data sets for these quality traits. In this regression approach we relate potato tuber quality traits to the obtained ~omics data sets within the C × E population. Each ~omics data set is treated as an independent predictor set and phenotypic traits as response variables. We validate the obtained results based on prior knowledge of regulatory and metabolic routes associated with flesh color. This approach resulted in associated genes and metabolites, some of which were already known to be involved in these traits and which confirm the validity of this approach. In addition, novel metabolites were found to be highly correlated to the phenotypic traits.

2. Materials and methods

2.1. Plant material

Ninety-six individuals, including the parental clones, of a diploid backcross population (C × E) [29] were used in this study. This population is derived from an original cross between potato clones C (USW533.7) and E (77.2102.37) and is described in detail in Celis-Gamboa et al. [17]. All clones were grown in multi-year repeats in the field, Wageningen, The Netherlands during the normal potato-growing season in the Netherlands (April–September). For each genotype, tubers were collected from three plants and representative samples were either used for phenotypic analysis or mechanically peeled and immediately frozen in liquid nitrogen before being ground into a fine powder and stored at −80 °C. Phenotypic data of potato flesh color and determination of carotenoids are described in Kloosterman et al. [16]. Enzymatic discoloration (ED) after 5 min, 30 min, 3 h and difference in discoloration between 3 h and 30 min were described in Werij et al. [28]. The Pearson correlation coefficient between enzymatic discoloration after 5 min and 30 min was very high (0.99), so out of these two we considered only enzymatic discoloration after 5 min.

2.2. Microarray hybridizations and data processing

RNA was extracted from the 96 samples using the hot phenol method described previously [15]. All samples were labeled with both Cy3 and Cy5-dye using the low RNA input linear Amplification Kit, PLUS, Two color (Agilent technologies) according to the manufacturer's protocol starting with 2 μg of purified total RNA. Hybridization and washing was performed according to the Agilent's two-color hybridization protocol with the following change: 1 μg of labeled Cy5 and Cy3 cRNA was used as input in the hybridization mixture. Slides were scanned on the Agilent DNA Microarray Scanner and data extracted using the feature extraction software package (v9.1.3.1) using a standard two-color protocol. Genes which show consistent low expression (<2 * BG) were removed and data sets were independently normalized using the quantile normalization procedure (mean) available in Genstat® 11.1. For additional data analyses only genes with a Pearson correlation coefficient higher than 0.8 between the Cy3 and Cy5 datasets were included resulting in 15,062 expressed genes. Expression data of associated genes can be found in Supplementary Table 1.

2.3. LC–MS data generation and data processing

Potato tuber samples were analyzed for variation in semi-polar metabolite composition using an untargeted accurate mass LC–MS approach, with on-line absorbance spectra measurements using a photodiode array (PDA) detector, essentially as described in De Vos et al. [32]. In short, 500 mg FW of frozen tuber powder was weighed in glass tubes and extracted with 1.5 ml of 87.5% methanol containing 0.125% formic acid. Samples were sonicated and centrifuged, and then filtered (Captiva 0.45 μm PTFE filter plate, Ansys Technologies) into 96-well plates with 700 μL glass inserts (Waters) using a TECAN Genesis Workstation. Extracts (5 μL) were injected using an Alliance 2795 HT instrument (Waters), separated on a Phenomenex Luna C18 (2) column (2.0 mm × 150 mm, 3 mm particle size) using a 45 min 5–35% acetonitrile gradient in water (both acidified with 0.1% formic acid) and then detected firstly by a photodiode array detector (Waters 2996) at a wavelength range of 220–600 nm and secondly by a Waters-Micromass QTOF Ultima MS with positive electrospray ionization at a mass range of m/z 100–1500. Leucine enkaphalin was used as lock mass for on-line mass calibration.

Metalign software (www.metalign.nl) was used to extract and align all accurate mass signals (with signal to noise ratio ≥3) from the raw data files. A total of 14,428 mass signals were thus obtained. Signals present in at least 10 samples and with at least one amplitude higher than 100 (about 5 times the noise value) were subsequently selected, resulting in a dataset of 3024 mass signals. Finally, the so-called multivariate mass spectra reconstruction strategy [34] was used to remove data redundancy by retention time-dependent clustering of signals derived from the same compound, *i.e.* isotopes, adducts and in-source fragments. This clustering of the 3024 signals revealed 233 reconstructed metabolites (centrotypes) and 425 (14%) single non-clustered mass signals. From each reconstructed metabolite (centrotypes) the signal intensity of the most unique mass was selected for further statistical analyses. The untargeted metabolites are represented as centrotype_mass_scan number. For a hypothetical example: 818_795_918 which means the centrotype number is 818, mass number 795 and scan number is 918.

Extraction and analyses of carotenoids, tocopherols and chlorophylls were performed as described by Kloosterman et al. [16]. In short, 500 mg of frozen powder was weighed and extracted twice with methanol–chloroform–Tris buffer containing 0.1% BHT as an antioxidant. The chloroform fractions were pooled, dried using nitrogen gas and taken up in 1 ml of ethylacetate. The chromatographic system consisted of a W600 pump system, a 996 PDA detector, and a 2475 fluorescence detector (Waters Chromatography). An YMC-Pack reverse-phase C30 column (250 mm × 4.6 mm, particle size 5 μm) at 40 °C was used to separate the compounds present in the extracts. Data were analyzed using Empower software (Waters Chromatography). Quantification of compounds was based on calibration curves constructed from respective standards carotenoids were extracted and analyzed by HPLC with photodiode array (PDA) detection [16]. For some of the carotenoids such as zeaxanthin (zea), violaxanthin (vio) and violaxanthin-like (vio-like) [16] we could not quantify the intensity values in a small subset of genotypes (for example: 21 genotypes for zeaxanthin, 20 genotypes for violaxanthin and violaxanthin-like) as they were below the detection limit. For statistical (regression) analysis and in order to avoid underestimation of the variability using a fixed threshold value we generated random data for the genotypes with values below the detection threshold using a uniform distribution between zero and the minimum value of the particular carotenoid (Supplementary Table 2). Results were not sensitive to this particular approach. In total we are considering 233 untargeted metabolomics centrotypes and combined these with three targeted metabolites: zeaxanthin (zea), violaxanthin(vio) and violaxanthin-like (vio-like). We used 86 genotypes (samples) which were common in LC–MS, transcriptomics (Cy3 and Cy5) and traits.

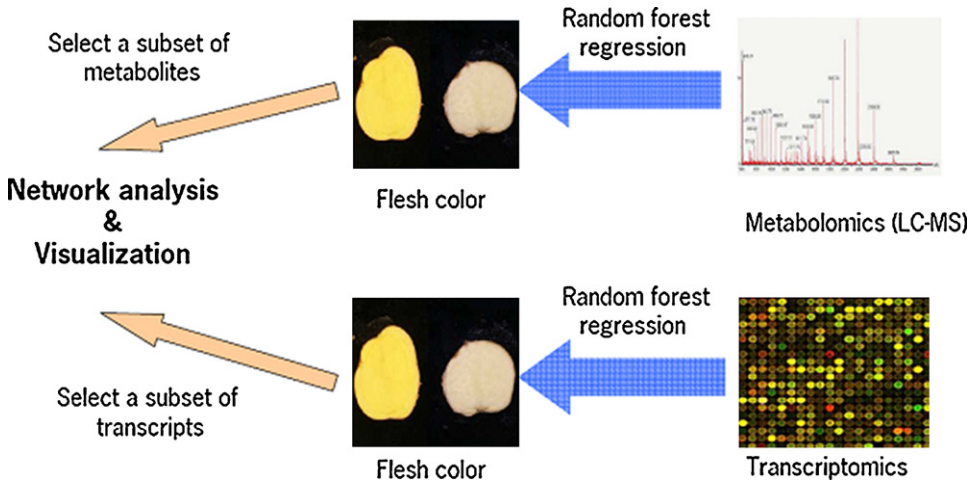

Fig. 1. Integration of metabolomics and transcriptomics data with a trait (potato flesh color or enzymatic discoloration) to identify important metabolites and genes associated with these traits, followed by a correlation network analysis and visualization.

2.3.1. Metabolite identification

Metabolites showing high correlation to phenotypic traits in the statistical analyses were (putatively) identified by comparing the detected accurate masses of the mono-isotopic molecular ions to those reported in the MotoDB (http://appliedbioinformatics.wur.nl/moto/) [33], the Dictionary of Natural Products (www.chemnetbase.com), the KNApSAcK database (http://kanaya.naist.jp/KNApSAcK) and/or the ChemSpider database (http://www.chemspider.com), using a mass deviation window of 5 ppm. Suggested elemental compositions and annotations of compounds were checked for the presence of corresponding in-source fragments within the mass clusters and for their UV/VIS absorbance spectra, if present, in the original raw data files.

2.4. Random Forest (RF) regression

A Random Forest [8,36] is a collection of unpruned decision trees [9], used mostly for statistical classification but this method can be applied for regression as well [10]. A Random Forest model is typically made up of hundreds of decision trees. Each decision tree is built from a bootstrap sample of the original data set. That is, some samples will be included more than once in a particular bootstrap sample, whereas others will not appear at all. Generally, about two thirds of the samples will be included in a bootstrap sample and one third will be left out (called the out-of-bag samples or OOB samples). The variance explained in RF ($R2$) is defined as 1 − (Mean square error (MSE)/Variance of response), where MSE is the sum of squared residuals of the OOB samples divided by the OOB sample size [39]. Since the MSE is estimated on the OOB samples and the total variance on all the samples, $R2$ can be negative. In each analysis, we estimated the variance explained by the RF model ($R2$) on the OOB samples, which is different from the $R2$ for goodness-of-fit in normal ordinary least square (OLS) regression [20]. Variance explained ($R2$) from RF is a value that is relevant for prediction of independent new samples, whereas the $R2$ in OLS is just a goodness-of-fit of the data at hand. Estimation of variable importance of the transcripts and metabolites was based on the Gini increase in MSE [8]. The greater the increase in the node purity values [8] the greater the importance of that particular variable [8]. We used for the "mtry" parameter one third of the total number of variables (metabolites or transcripts) used. So, for metabolomics data set, we used 78 (236/3) and for transcriptomics data set, the value was 5020 (15,062/3).

We used Random Forest (RF) for regression of the phenotypic traits flesh color and enzymatic discoloration at the different time points individually and separately for the transcriptomics Cy3, transcriptomics Cy5 and the metabolomics data sets. All three data sets were $\log_2$ transformed and then autoscaled (mean = 0, sd = 1).

2.5. Permutation test for statistical significance

RF quantifies the importance of genes and metabolites that explain the variation present in flesh color and enzymatic discoloration (ED), but does not give a significance level or a threshold to choose a possible subset of associated genes or metabolites. Therefore, we included a permutation test to indicate significance of each gene and metabolite association in this study. In our situation, we randomized all phenotypic traits separately (for example: flesh color) and each time applied RF, separately for the transcriptomics and metabolomics data set. The RF model was applied 1000 times for 1000 different randomizations of the trait values and in each analysis we estimated the variance explained by the RF model ($R2$) and variable importances of all variables in terms of decrease in node impurities. We ordered node purity values from the permuted data sets and took the 95 percentile from the distribution impurity values for node impurity to assess the significance of the of individual genes and metabolites. The same was done for $R2$ values of the model: the 95-percentile was used as a cutoff to denote significance of an $R2$ value in RF regression.

RF regression of flesh color and enzymatic discoloration on gene expression values (Cy3 and Cy5 intensities) and metabolites separately were conducted using the "Random Forest" package of R statistical software.

For transcriptomics (both for Cy3 and Cy5) and metabolomics analysis with flesh color and enzymatic discoloration we took into account the top 50 significant filtered genes and metabolites to validate our data integration approach. We mapped expression QTLs (eQTLs) from the gene expression data and metabolite QTLs (mQTLs) from the LC–MS data to find loci explaining genetic variation in metabolites and gene expression values using the Metanetwork package of R statistical software [25]. Separate linkage maps for "C" (Cmap) and "E" (Emap) parent were used [38] for QTL analysis.

2.6. Network reconstruction

A network [26] is a set of nodes (vertices) and a set of edges. Nodes represent either genes, metabolites or a trait whereas edges represent associations. Pearson correlation coefficients were used to quantify the strength of association between all combinations of genes, metabolites and phenotypic traits. The significance thresh-

Table 1
Associated transcripts and eQTL analysis after Random Forest regression for flesh color and enzymatic discoloration.

Gene ID	Rank of genes	Expression QTL		
	(Cy5,Cy3)	Cmap	Emap	Description blastX
Flesh color				
MICRO.7880.C2	1,1	3	–	Beta carotene hydroxylase [Lycopersicon esculentum]
MICRO.1510.C2	2,2	3	–	Salt tolerance protein 5-like protein [Solanum tuberosum]
MICRO.7880.C1	3,5	3	–	Beta carotene hydroxylase [Lycopersicon esculentum]
STDB005D11u.scf	4,4	3	–	Hairpin-induced family protein [Ipomoea nil]
STMIT26TV	5,3	3	–	NA
ACDA00891C03.T3m.scf	6,8	3	–	Putative Kunitz-type tuber invertase inhibitor precursor [Solanum tuberosum]
MICRO.7862.C1	7,6	3	–	Os11g0206700 [Oryza sativa (japonica cultivar-group)]
STMHD34TV	8,9	3	–	NA
MICRO.15198.C2	9,19	3	–	Os09g0363500 [Oryza sativa (japonica cultivar-group)]
bf_mxflxxxx_0066c08.t3m.scf	10,7	3	–	NA
MICRO.6275.C2	11,11	3	–	Hypothetical protein [Plantago major]
MICRO.11569.C1	12,12	3	–	Heat shock protein 82
MICRO.16733.C1	13,16	3	–	Hypothetical protein [Phaseolus vulgaris]
MICRO.9632.C3	14,18	2	–	Unknown protein [Arabidopsis thaliana]
MICRO.4880.C1	16,13	3	3	Os02g0773300 [Oryza sativa (japonica cultivar-group)]
MICRO.16246.C1	17,24	2	2	ATEB1C (MICROTUBULE END BINDING PROTEIN 1); microtubule binding [Arabidopsis thaliana]
MICRO.17254.C1	18,30	2	2	NA
MICRO.10804.C1	19,10	3	–	Hydrolase/zinc ion binding [Arabidopsis thaliana]
SDBN006M13u.scf	20,15	3	–	Cytochrome P450 71A8
bf_mxlfxxxx_0066e10.t3m.scf	22,14	3	–	NA
MICRO.14821.C1	23,23	2	2	NA
MICRO.7742.C2	25,20	3	–	Unknown protein [Arabidopsis thaliana]
MICRO.13887.C1	26,44	2	2	Zeaxanthin epoxidase [Solanum tuberosum]
MICRO.729.C1	27,42	3	3	Unknown protein [Arabidopsis thaliana]
MICRO.14225.C2	30,26	3	–	CONSTANS interacting protein 2a [Lycopersicon esculentum]
MICRO.12704.C1	31,17	–	2	Beta tubulin; Remorin, C-terminal region [Medicago truncatula]
MICRO.17262.C1	32,41	2	2	Hypothetical protein [Cleome spinosa]
MICRO.9413.C1	33,38	3	–	ETEA-like (expressed in T-cells and eosinophils in atopic dermatitis) protein [Brachypodium sylvaticum]
bf_arrayxxx_0102h05.t7m.scf	35,28	2	–	Lipase, class 3 [Medicago truncatula]
MICRO.14714.C1	38,25	3	3	Orcinol O-methyltransferase [Rosa hybrid cultivar]
bf_mxflxxxx_0035e07.t3m.scf	39,34	3	–	Cytochrome P450 71A2 (CYPLXXIA2) (P-450EG4)
bf_mxlfxxxx_0025b07.t3m.scf	41,37	2	2	Os12g0582800 [Oryza sativa (japonica cultivar-group)]
MICRO.3489.C1	44,27	3	–	Os05g0572700 [Oryza sativa (japonica cultivar-group)]
bf_ivrootxx_0062f07.t3m.scf	46,46	2	2	Ran GTPase binding/chromatin binding [Arabidopsis thaliana]
MICRO.15175.C1	48,48	3	–	Putative beta 1,3 glucan synthase [Oryza sativa (japonica cultivar-group)]
MICRO.729.C1	1,4	3	3	Unknown protein [Arabidopsis thaliana]
POACL49TP	2,2	8	4	cAMP response element binding (CREB) protein [Medicago truncatula]
MICRO.14874.C1	3,20	8	4	cAMP response element binding (CREB) protein [Medicago truncatula]
MICRO.16330.C1	4,1	3	1	Os03g0717600 [Oryza sativa (japonica cultivar-group)]
Enzymatic discoloration				
MICRO.8173.C2	5,3	3	3	Unknown protein [Arabidopsis thaliana]
SSBT003N21x.scf	7,37	–	1	NA
MICRO.10002.C2	8,28	3	3	Sialyltransferase-like protein [Glycine max]
MICRO.16733.C1	17,8	3	–	Hypothetical protein [Phaseolus vulgaris]
SSBT005M24x.scf	19,48	–	–	NA
bf_mxflxxxx_0006b11.t3m.scf	25,10	3	3	NA
MICRO.1371.C1	44,36	8	–	ATP binding/shikimate kinase [Arabidopsis thaliana]
MICRO.7678.C1	50,41	8	8	Unknown protein [Arabidopsis thaliana]

old ($\alpha = 0.01$) was used to draw lines between genes, metabolites or traits. Only significant relationships ($p < 0.01$) are drawn.

2.7. Software

Statistical analyses were performed in the "R" statistical programme (http://www.r-project.org/) using the "randomForest" package. For network visualization we used Pajek software [27].

The schematic diagram of the methodology is shown in Fig. 1.

3. Results

3.1. RF regression on the transcriptomics data set

3.1.1. Flesh color

Random Forest regression was applied to the transcriptomics data set using flesh color as the response. The variance in flesh color explained by the RF using Cy3 gene expression ($R2$) was 58%. In the top 50 of associated genes ranked by their variable importance a beta-carotene hydroxylase gene (Bch) ranks first and another copy of the same gene ranks third position, while another gene in the carotenoid pathway, zeaxanthin epoxidase (Zep), was ranked forty-fourth (Supplementary Table 3). Based on our current knowledge of the potato tuber flesh color or carotenoid content [16] these genes were expected to be associated with flesh color.

The RF model explained 58% (Table 3) of the variance and was significant at permutation threshold of 0.001 and 233 genes were found to be significantly associated (at permutation p-value <0.001). The variance explained by 233 significant genes was 73%. The technical repeat of the transcriptomics data set using a second labeling dye (Cy5) was analyzed using the same approach as for the Cy3 data set and here the RF model explained 60% of the variance in flesh color. In the top 50 of genes ranked by variable importance, beta-carotene hydroxylase (Bch) ranks first again, while zeaxan-

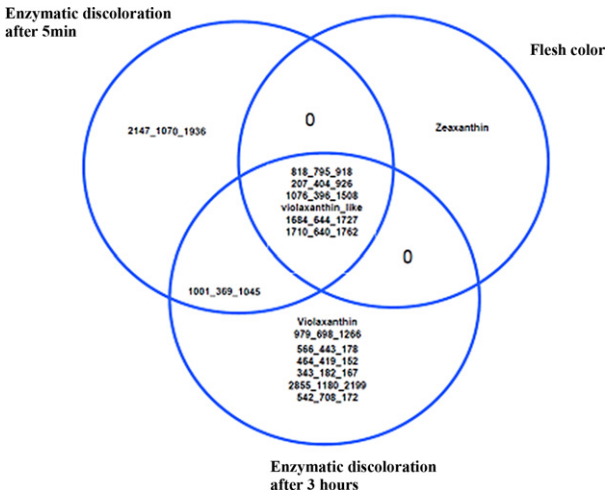

Fig. 2. Venn diagram of metabolites associated with flesh color, enzymatic discoloration after 5 min. Six metabolites are in common.

thin epoxidase (Zep), now ranks twenty-second. Similar to the Cy3 data set, 304 genes are found significant at a permutation *p*-value threshold of 0.001. The *R*2 value (60%) was also significant at the permutation threshold of 0.001.The variance explained by 304 significant genes was 75%. All genes significant in the analysis of Cy3 data were also found significant for the Cy5 data. Between the top 50 sets of genes of the Cy3 and Cy5 data, we found 35 genes in common (see Table 1, first column). Using only those 35 significant filtered genes in Cy3 and Cy5 the variance explained by the RF model was 71% for Cy3 and 72% for Cy5 data sets.

3.1.2. Enzymatic discoloration

Enzymatic discoloration measured at different time points was also regressed on the transcriptomics data sets (Cy3 and Cy5, separately). Only enzymatic discoloration after 5 min had a positive value for the variance explained by the RF model (Table 3).

For the Cy3 data set, 420 genes were significant at the permutation significance threshold level of 0.001. The variance explained (14%) was also significant at this level. The variance explained by 420 significant genes was 51%. For the Cy5 data, 438 genes are significant (permutation $p < 0.001$). The amount of variance explained by the RF model (17%) was also significant at this level. The variance explained by only 438 significant genes was 47%.

For enzymatic discoloration at the other time points the RF model was not significant and we actually obtained negative values for the variance explained by the model (this is possible because the *R*2 value is obtained from the OOB samples).

Between the Cy3 and Cy5 data sets, 12 genes were in common in the top 50 genes (Table 1). Interestingly between flesh color and enzymatic discoloration after 5 min, 2 genes (Gene IDs: MICRO.16733.C1 and MICRO.729.C1) were in common (shown in bold, Table 1). The variance explained by a RF model using only the 12 genes in Cy3 and Cy5 was 48% and 46%, respectively.

3.2. RF regression on the metabolomics data set

3.2.1. Flesh color

We also applied RF to the metabolomics data set obtained from the same material used for expression analysis. The variance of flesh color explained by the RF model for metabolites was 63%. Seven metabolites were significant (permutation threshold 0.001) (Fig. 2), the *R*2 value was also significant at this level. Using only these seven metabolites the variance explained by the RF model was 77%.

3.2.2. Enzymatic discoloration

Enzymatic discoloration after 5 min, 3 h and difference in discoloration between 30 min and 3 h, were regressed (separately) on the metabolomics data set. The variance explained by the metabolites in enzymatic discoloration after 5 min, 3 h and difference in discoloration between 3 h and 30 min was 16%, 10% and 1%, respectively. The model for enzymatic discoloration after 5 min and after 3 h is just significant at 0.001 level. Eight metabolites were significant for enzymatic discoloration after 5 min, 14 for enzymatic discoloration after 3 h and seven of these overlap between both data sets. Six of these are also in common with flesh color (Fig. 2)

3.3. Integration of transcriptomics and metabolomics data

For flesh color, we used the 35 significant genes in common in the top 50 s of the Cy3 and Cy5 data sets and the seven metabolites from the LC–MS data set. We combined these genes and metabolites in a single data set and applied RF regression. The variation explained by the combination of Cy3 and metabolomics data sets was 82%. This is an increase by 11% and 5% for the independent data sets, transcriptomics (Cy3) and metabolomics, respectively. The variation explained by the combination of Cy5 and metabolomics integrated data sets was 78%. This is an increase by 6% and 1% for the independent data sets, transcriptomics (Cy5) and metabolomics, respectively. If we integrate the Cy3 and metabolites whole data sets (and not just the significant ones) then the variance explained by the model is 64%. In case of Cy5 and the metabolites, 62% of the variance is explained by the model.

For enzymatic discoloration (5 min) we used the 12 significant genes in common between the Cy3 and Cy5 data sets and the eight metabolites from metabolomics RF analysis. After combining Cy3 and metabolomics data into a single data matrix, we applied RF regression and the explained variance now was 50%. We see an increase in explained variance in the integrated data in comparison to the individual data sets by 2% and 26% for the transcriptomics (Cy3) and metabolomics data sets, respectively. For Cy5 and the metabolomics data set, the variance explained by the RF model was 51%. Again, we see an increase in explained variance in the integrated data in comparison to the individual data sets by 4% and 27% for transcriptomics (Cy5) and metabolomics, respectively. Also here if we integrate Cy3 and metabolites whole datasets (and not just significant ones) then the variance explained by the model is 14%. In case of Cy5 and metabolites, 15% of the variance is explained by the model.

For additional enzymatic discoloration data (3 h and difference 3 h-30 min), we did not get any significant RF regression model (*R*2 values were negative) even if we take the entire gene set (Cy3 or Cy5) with the metabolites.

3.4. Network reconstruction

For network reconstruction of potato tuber flesh color, we used the seven associated metabolites as well as two targeted metabolites zeaxanthin and violaxanthin_like and the two genes Bch and Zep. Six of the metabolites significant for enzymatic discoloration were in common with flesh color (Fig. 2).

For flesh color, we took two significant genes (Bch, Zep) from the transcriptomics data set which were already reported in the literature [18,19] as being involved in the carotenoid pathway and also found in transcriptomics analysis in the top 50 genes for Cy3 and Cy5 datasets separately and significant mass scans from the metabolomics study. Hence for network reconstruction (Fig. 3) we took two genes and significant metabolites.

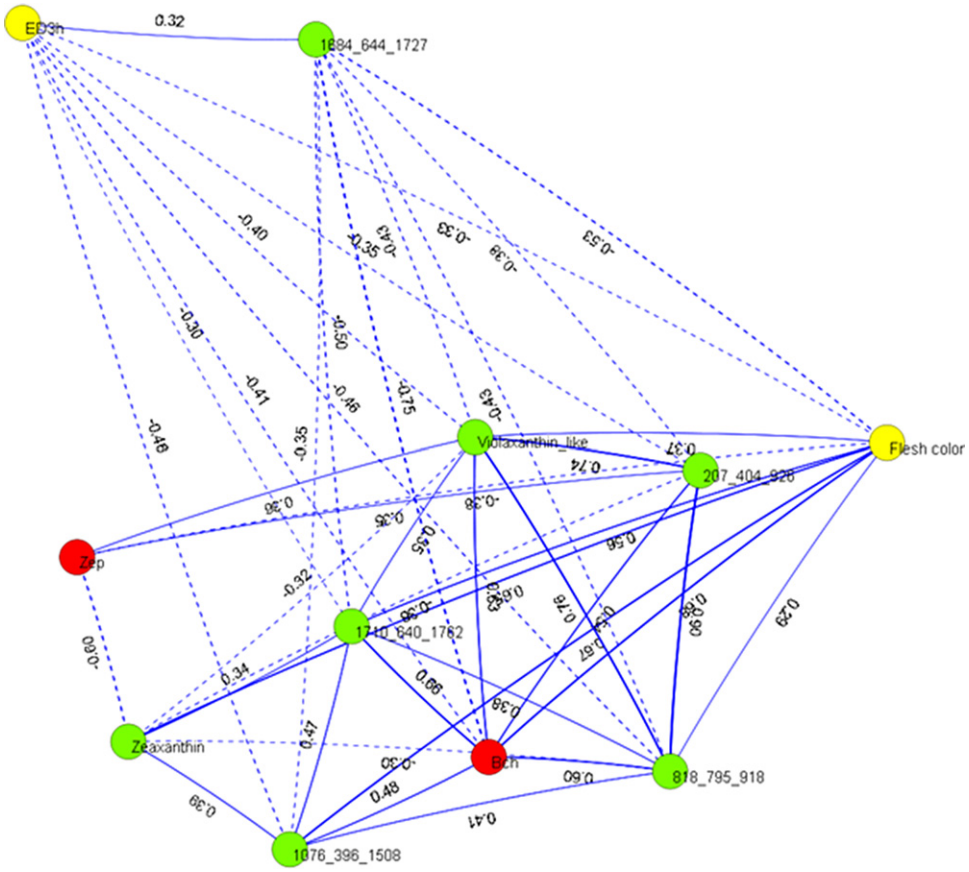

Fig. 3. Pearson correlation network of genes (red), metabolites (green) and phenotypic traits (yellow). The dotted lines represent negative correlation coefficients, solid lines represent positive correlation coefficients. Only correlation coefficients significant at $p < 0.01$ are considered. Bch = beta-carotene hydroxylase, Zep = zeaxanthin epoxidase, ED3h = enzymatic discoloration after 3 h.

4. Discussion and conclusions

4.1. Flesh color

We used RF regression for integrating transcriptomics and metabolomics data sets with potato tuber flesh color. In the transcriptomics analysis, 35 genes were in common between the top 50 associated genes of the Cy3 and Cy5 data sets. In that gene set beta-carotene hydroxylase (Bch) [18] and zeaxanthin epoxidase (Zep) [19] were ranked high, both of which have previously been associated with flesh color in potato tubers [18,19]. Bch catalyzes a crucial step in the biosynthesis of carotenoids and a dominant allele of Bch (B) exhibits higher expression resulting in an expected increase of carotenoid biosynthesis through conversion of B-carotene to zeaxanthin [16]. Zeaxanthin epoxidase (Zep) catalyzes the first step in the conversion of zeaxanthin to violaxanthin. The presence of a recessive Zep allele in homozygous state results in the accumulation of zeaxanthin in the tuber [19]. However, in order for zeaxanthin to accumulate to relatively high levels, the dominant allele of Bch is required, thus presenting a non-linear relation which was successfully detected with the RF regression approach. A similar approach was used by Jiang et al. [35] used for identification of epistatic interactions.

The majority of the additional genes with high ranking show eQTLs on chromosomes 3 and 2 directly under the QTLs for flesh color. If one considers the flesh color QTL on chromosome 3 as a single locus QTL with Bch as the responsible gene, then all other genes with map positions on chromosome 3 can be either considered false positive due to genetic linkage with the genome region or the variation in gene expression is directly linked to Bch activity and thus carotenoid levels. A similar reasoning can be applied to the QTL on chromosome 2 where Zep is considered to be responsible for the accumulation of zeaxanthin. Other genes may also have a functional relationship to the carotenoid pathway, flesh color or enzymatic discoloration but based on their annotation and the possibility of significance due to linkage, we consider that a less likely explanation. With the publication of the first draft of the potato genome sequence we can now assess the physical position of genes within the genome in relation to the position of the QTLs. If we check the physical positions of the 35 genes associated with flesh color we find a large representation of genes (24 genes) on chromosomes 3 and 2 in the same regions as the flesh color QTL support intervals suggesting genetic linkage to the respective genomic regions and not functional association (data not shown). Interestingly, a gene with homology to a orcinol O-methyltransferase, involved in flavonoid biosynthesis [41], has an eQTL coinciding with the QTL for flesh color on chromosome 3 (Table 1) but the gene itself resides on chromosome 6, indicating trans-acting transcriptional control. This association would imply a novel interesting link between flavonoid and carotenoid metabolism, however, as there is not yet any biological supporting evidence for such a direct relation it was not included in the network analysis. Interesting to note is the absence of other characterized carotenoid biosynthesis genes or carotenoid cleavage de-oxygenases (CCDs) involved in carotenoid breakdown [30]. In our analysis, associations are largely dependent on variation in gene expression, whereas for many traits variation in enzyme activities determine the phenotype.

From the metabolomics analysis we obtained seven metabolites which are significantly associated with potato flesh color. Based on their accurate masses and in-source fragmentation

Table 2
Associated metabolites and mQTL analysis after RF regression for flesh color and enzymatic disocoloration.

Metabolites	Metabolite QTL		
	Cmap	Emap	Description
Flesh color			
818_795_918	3	–	2,3-Dihydroxy-4-megastigmen-9-one-glucoside
1076_396_1508	3	–	4,7-Megastigmadiene-3,9-diol-glucoside
207_404_926	3	2	–
1684_644_1727	3	3	–
1710_640_1762	3	–	–
Violaxanthin_like	3	–	Violaxanthin_like
Zeaxanthin	2	–	Zeaxanthin
Enzymatic discoloration			
1001_369_1045	–	2	Caffeoylquinic acid methyl ester
343_182_167	–	5	Tyrosine
Violaxanthin	–	–	Violaxanthin
979_698_1266	9	1	–
566_443_178	–	–	–
464_419_152	–	5	–
2147_1070_1936	–	–	–
2855_1180_2199	–	5	–

Table 3
Variance explained ($R2$), significance of the RF model and numbers of significant variables ($\alpha = 0.001$).

Name of the traits	Cy3	Cy5	LC peaks
Flesh color	58 (233)[a]	60 (304)[a]	63 (7)[a]
Enzymatic discoloration after 5 min	14 (420)[a]	17 (438)[a]	16 (8)[a]
Enzymatic discoloration after 3 h	Negative	Negative	10 (14)[a]
Enzymatic discoloration difference between 3 h and 30 min	Negative	Negative	1[b]

[a] Number of significant genes (Cy3/Cy5) or LC peaks at 0.001 significant level.
[b] Model is not significant at 0.001.

indicating the presence of a hexose unit, probably glucose, mass peaks 1076_396_1508 (centrotype 1076) and 818_795_918 (centrotype 818) were putatively identified as 4,7-megastigmadiene-3,9-diol-glucoside and 2,3-dihydroxy-4-megastigmen-9-one-glucoside, respectively. These compounds are non-volatile glycosides of carotenoid-derived volatile metabolites. Based on the negative and positive correlation of centrotype 818 with, respectively, zeaxanthin and violaxanthin-like (Fig. 3), one could infer 818 is synthesized downstream of zeaxanthin. In contrast, centrotype 1076 is positively correlated with zeaxanthin with no significant correlation to violaxanthin or violaxanthin-like, indicating this metabolite derives directly or indirectly from zeaxanthin. Hence such network reconstruction from the transcriptomics and metabolomics data sets help us to build a hypothesis [23] regarding possible existence and position of the genes or metabolites in metabolic pathways; however, some prior knowledge regarding genes or metabolites is required to reconstruct such network. To be more certain about the annotation of these putative carotenoid-derived metabolites, further chemical analyses, such as hydrolysis followed by GC–MS of the volatile compounds and preferably NMR, are needed.

4.2. Enzymatic discoloration

In the transcriptomics analysis only enzymatic discoloration after 5 min was significantly associated with gene expression. Twelve genes were in common between the top 50 genes of the Cy3 and Cy5 data sets. In the eQTL analysis of those genes, four genes were mapped to chromosome 8, five genes to chromosome 3, one gene to chromosome 1 and two genes to chromosome 5. In a previous QTL study [28], QTLs for enzymatic discoloration were mapped to chromosomes 1, 3 and 8. Polyphenol oxidase (PPO) is considered the candidate gene for the QTL on chromosome 8 by catalyzing the oxidation of phenolic compounds. Candidate genes associated with the remaining QTLs have not yet been identified. From the metabolomics analysis, out of 8 significant untargeted metabolites we putatively identified two metabolites as Caffeoylquinic acid methyl ester and tyrosine, which were already reported by Werij et al. [28] and associated with enzymatic discoloration and confirm our findings. Other metabolites in Table 2 (such as 2147_1070_1936, 979_698_1266) we could not identify yet (due to too low abundance or multiple charged ions) even though they showed significant association in the RF analysis.

Interestingly there seems to be a strong overlap between metabolites associated with flesh color and enzymatic discoloration (Fig. 2). QTL analysis of tuber flesh color and enzymatic discoloration shows a similar overlap: a QTL on chromosome 3 due to multiple QTL effect.

In apricot the level of certain carotenoids has been associated with inhibition of enzymatic browning reactions through inhibition of PPO activity and substrate regeneration [31]. The question arises whether variation in the amount of carotenoids is partially responsible for varying degrees of enzymatic discoloration and secondly what metabolites/carotenoids are involved in potato tubers.

For potato flesh color, the variation explained by the RF model using transcriptomics data was 58% (Cy3) and 60% (Cy5) whereas using only the metabolomics data, 63% was explained. The variation explained by the RF model with filtered 35 significant genes was 71% for Cy3 and 72% for Cy5 data set whereas with seven metabolites was 77%. The combination of all three data sets gives 82% (Cy3 and metabolomics data sets) and 78% (Cy5 and metabolomics data) of explained variance.

Similarly, for enzymatic discoloration after 5 min, the variation explained by the RF model using transcriptomics data was 14% (Cy3) and 17% (Cy5) whereas using only metabolomics data this was 16%. The variation explained by the RF model with filtered 35 significant genes was 48% for Cy3 and 46% for Cy5 data set where as with eight metabolites was 24%. The combination of all three data sets gives 50% (Cy3 and metabolomics data) and 51% (Cy5 and metabolomics data sets) explained variance.

From this analysis we observe that the improvement of the RF model is because of two reasons: first, due to the filtering out of significant genes or metabolites from the individual data sets. Thus we get rid of the variables which are not associated with the trait (noise variables). Although we get hundreds of significant genes, we take into account only the top 50 genes because the associations with the trait with the top 50 are stronger than with the rest of the significant genes. Although such choice is arbitrary but it helps in reducing the numbers of significant genes and validate our approach with previously reported genes linked with flesh color pathway such as "Bch" and "Zep". The second improvement is due to the integration of the filtered genes and metabolites into a single data set.

For enzymatic discoloration, the variance explained by the model is much lower than for flesh color, which might be due to the higher difficulty in phenotypic scoring of discoloration. Some other likely explanations are errors in the measuring procedure, possibly higher influence of environmental effects on discoloration than on flesh color per se, and the involvement of more genes in enzymatic discoloration than in flesh color.

We used RF regression for integrating the transcriptomics (Cy3 and Cy5) and metabolomics data sets with phenotypes of interest. This procedure can handle high dimensional data (for example, number of genes are much higher than the number of samples) and has an internal cross-validation [8] (using the OOB samples) and has only a few tuning parameters which if chosen reasonably, do not change results strongly [40]. It could be considered a limitation that a RF model by default will (or, at least, can) use all variables simultaneously and if we want to perform variable selection, we need to set a threshold on the number of variables or we need to select variables based on a significance criterion or a variable selection procedure.

In this paper we used genetic information through QTL analysis on the one hand and prediction of the traits using RF analysis from transcriptomics and metabolomics analysis on the other hand. From QTL analysis, we can identify the map position of the genes or metabolites but due to linkage we also get false positives.

In RF regression approach prediction of phenotype from metabolomics or transcriptomics data is possible in a way that genes and metabolites might be linked with phenotype but independent of the genetic information [42]. Combining both QTL analysis and prediction of traits using RF gives us a clue about the candidate genes and metabolites which are linked with phenotype but also the genetic information about those genes and metabolites from QTL analysis.

We have applied RF regression as a tool for data integration of metabolites and gene expression profiles relating to a phenotypic trait of interest, but the same methodology can be used for data integration with a quantitative variable and other ∼omics data sets such as transcriptomics and proteomics, metabolomics and proteomics or metabolomics, transcriptomics and proteomics data. Where it can help to identify and hypothesize the components (genes, metabolites, proteins, etc.) in a pathway of interest and the genetic basis of the genes, metabolites or proteins involved in the pathway.

Acknowledgement

This work is supported by the Wageningen UR Plant Breeding, Wageningen, The Netherlands.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.aca.2011.03.050.

References

[1] A. Brazma, J. Vilo, FEBS J. 480 (2000) 17–24.
[2] T. Gaasterland, S. Bekiranov, Nat. Genet. 24 (2000) 204–206.
[3] O. Fiehn, Plant Mol. Biol. 48 (2002) 155–157.
[4] W.B. Dunn, N.J.C. Bailey, H.E. Johnson, Analyst 130 (2005) 606–625.
[5] R. Aebersold, M. Mann, Nature 422 (2003) 198–207.
[6] H. Zhu, M. Bilgin, M. Snyder, Annu. Rev. Biochem. 72 (2003) 783–812.
[7] A.R. Joyce, B. Palsson, Nat. Rev. Mol. Cell Biol. 7 (2006) 198–210.
[8] L. Breiman, Random forests, Mach. Learn. 1 (2001) 5–32.
[9] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer-Verlag, New York, 2001.
[10] M.R. Segal, Center for Bioinformatics & Molecular Biostatistics, University of California, San Francisco, 2004.
[11] H. Kitano, Science 295 (2002) 1662–1664.
[12] H. Kitano, Nature 420 (2010) 206–210.
[13] T.Y. Kim, H.U. Kim, S.Y. Lee, Curr. Opin. Biotechnol. 21 (2010) 78–84.
[14] A. Fukushima, M. Kusano, H. Redestig, M. Arita, K. Saito, Curr. Opin. Chem. Biol. 13 (2009) 532–538.
[15] C.W. Bachem, R.S. van der Hoeven, S.M. de Bruijn, D. Vreugdenhil, M. Zabeau, R.G.F. Visser, Plant J. 5 (1996) 745–753.
[16] B. Kloosterman, M. Oortwijn, J. UitdeWilligen, T. America, R. de Vos, R.G.F. Visser, C.W.B. Bachem, BMC Genomics 11 (2010) 158.
[17] B.C. Celis-Gamboa, P.C. Struik, E. Jacobsen, R.G.F. Visser, Ann. Appl. Biol. 143 (2003) 175–186.
[18] C.R. Brown, T.S. Kim, Z. Ganga, K. Haynes, D. De Jong, M. Jahn, I. Paran, W. De Jong, Am. J. Potato Res. 83 (2006) 365–372.
[19] A.M.A. Wolters, J.G.A.M.L. Uitdewilligen, B.A. Kloosterman, R.C.B. Hutten, R.G.F. Visser, H.J. Van Eck, Plant Mol. Biol. 73 (2010) 659–671.
[20] C.D. Montgomery, E.A. Peck, Introduction to Linear Regression Analysis, Wiley, New York, 1992.
[21] M. Bylesjo, D. Eriksson, M. Kusano, T. Moritz, J. Trygg, Plant J. 52 (2007) 1181–1191.
[22] S. Wienkoop, K. Morgenthal, F. Wolschin, M. Scholz, J. Selbig, W. Weckwerth, Mol. Cell. Proteomics 7 (2008) 1725–1736.
[23] A. Zamboni, M.D. Carli, F. Guzzo, M. Stocchero, S. Zenoni, A. Ferrarini, P. Tononi, K. Toffali, A. Desiderio, K.S. Lilley, M.E. Pè, E. Benvenuto, M. Delledonne, M. Pezzotti, Plant Physiol. 154 (2010) 1439–1459.
[24] N. Delmotte, C.H. Ahrens, C. Knief, E. Qeli, M. Koch, H.M. Fischer, J.A. Vorholt, H. Hennecke, G. Pessi, Proteomics 10 (2010) 1391–1400.
[25] J. Fu, M.A. Swertz, J.J. Keurentjes, R.C. Jansen, Nat. Protoc. 2 (2007) 685–694.
[26] J.S. Yuan, D.W. Galbraith, S.Y. Dai, P. Griffin, C.N. Stewart Jr., Trends Plant Sci. 13 (2008) 165–171.
[27] V. Batagelj, A. Mrvar, Pajek – Analysis and Visualization of Large Networks Graph Drawing Software, Springer, Berlin, 2003.
[28] J.S. Werij, B. Kloosterman, C. Celis-Gamboa, C.H. Ric de Vos, T. America, R.G.F. Visser, C.W.B. Bachem, Theor. Appl. Genet. 115 (2007) 245–252.
[29] B.C. Celis-Gamboa, Ph.D. Thesis, Wageningen University, Wageningen, The Netherlands, 2002.
[30] R. Campbell, L.J.M. Ducreux, W.L. Morris, J.A. Morris, J.C. Suttle, G. Ramsay, G.J. Bryan, P.E. Hedley, M.A. Taylor, Plant Physiol. 154 (2010) 656–664.
[31] D. Rigal, F. Gauillard, F.R. Forget, J. Sci. Food Agric. 80 (2000) 763–768.
[32] R.C.H. De Vos, S. Moco, A. Lommen, J.J. Keurentjes, R.J. Bino, R.D. Hall, Nat. Protoc. 2 (2007) 778–791.
[33] S. Moco, R.J. Bino, O. Vorst, H.A. Verhoeven, J. de Groot, T.A. van Beek, J. Vervoort, C.H. De Vos, Plant Physiol. 141 (2006) 1205–1218.
[34] Y. Tikunov, A. Lommen, C.H.R. De Vos, H.A. Verhoeven, R.J. Bino, R.D. Hall, A.G. Bovy, Plant Physiol. 139 (2005) 1125–1137.
[35] R. Jiang, W. Tang, X. Wu, W. Fu, BMC Bioinformatics 10 (2009).
[36] R. Díaz-Uriarte, S. Alvarez de Andrés, BMC Bioinformatics 7 (2006) 3.
[37] B.C.Y. Collard, M.Z.Z. Jahufer, J.B. Brouwer, E.C.K. Pang, Euphytica 142 (2005) 169–196.
[38] M.W. Bonierbale, R.L. Plaisted, S.D. Tanksley, Genetics 120 (1988) 1095–1103.
[39] H. Pang, A. Lin, M. Holford, B.E. Enerson, B. Lu, M.P. Lawton, E. Floyd, H. Zhao, Bioinformatics 16 (2006) 2028–2036.
[40] P.O. Gislason, J.A. Benediktsson, J.R. Sveinsson, Pattern Recogn. Lett. 4 (2006) 294–300.
[41] C. Stushnoff, L.J.M. Ducreux, R.D. Hancock, P.E. Hedley, D.G. Holm, G.J. McDougal, J.W. McNicol, J. Morris, W.L. Morris, J.A. Sungurtas, S.R. Verrall, T. Zuber, M.A. Taylor, J. Exp. Bot. 61 (2010) 1225–1238.
[42] M. Steinfath, N. Strehmel, R. Peters, N. Schauer, D. Groth, J. Hummel, M. Steup, J. Selbig, J. Kopka, P. Geigenberger, J.T. van Dongen, Plant Biotech. J. 8 (2010) 900–911.