Systems biology and statistical data integration of ~omics data sets

Animesh Acharjee

Thesis committee

Promotor
Prof. Dr. R.G.F. Visser
Professor of Plant Breeding
Wageningen University

Co-promotor
Dr. Ir. C.A. Maliepaard
Assistant professor at the Laboratory of Plant Breeding
Wageningen University

Other members
Prof. Dr. A.K. Smilde, University of Amsterdam
Prof. Dr. F.A. van Eeuwijk, Wageningen University
Prof. Dr. H.J. Bouwmeester, Wageningen University
Prof. Dr. Ir. M. Koornneef, Wageningen University and Max Planck Institute for Plant Breeding Research, Köln, Germany

This research was conducted under the auspices of the Graduate School Experimental Plant Sciences

Systems biology and statistical data integration of ~omics data sets

Animesh Acharjee

Thesis
submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus
Prof. dr. M.J. Kropff,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Wednesday 12 June 2013
at 1.30 p.m. in the Aula.

Animesh Acharjee
Systems biology and statistical data integration of ~omics data sets
177 pages.

PhD thesis, Wageningen University, Wageningen, The Netherlands, 2013
With references, with summaries in Dutch and English

ISBN 978-94-6173-584-3

Contents

Chapter 1

General introduction

Potato breeding and potato quality traits

Potato (*Solanum tuberosum L.*) is the world's third major food crop in terms of food consumption, and number eight in terms of area under cultivation (FAO statistics 2008). The potato tuber is a high-energy staple food in many countries around the world.

Today's world has a high demand of potatoes for consumption and industrial applications. To meet this demand, a continuous improvement of the potato crop by breeding is required. One of the major goals in potato breeding is high yield and an improved agronomic performance. An improvement in agronomic performance can be made in different areas such as introduction of resistance to pests and diseases, tolerance to abiotic stress factors like salinity, drought or low mineral content of soils, physiological and plant architectural traits leading to improved agricultural performance. Another area of improvement comprises quality aspects and it is an area that is likely to become ever more important in the near future. Depending on the target market, focus is shifting more towards breeding for quality. For instance, from a practical grower's perspective, being able to produce outstanding potato quality can offer a distinctiveness that can create extra revenue (Caswell and Mojduszka 1996). The physical properties connected to quality traits, like shape, colour and size, are easily observable by consumers and are therefore major factors determining market value of a product.

The nutritional composition of potatoes is important considering the high use of this crop in diets of people around the world. Potato is rich in carbohydrate content but it also provides significant quantities of other nutrients such as proteins, minerals and vitamins (Kadam et al., 1991). For the industry (e.g. crisps) quality in terms of texture, colour, frying properties, composition in terms of starch and sugars is very important as well.

Potato skin colour, eye depth, size and shape are crucial quality aspects of fresh potatoes for consumers, as they are immediately obvious while making the purchase. Therefore, these as well as other quality traits, are being considered in breeding and genetic research (Werij 2011). In cultivated potato, the flesh colour is predominantly white or yellow. Strong cultural preference exists for either white, preferred in the US and UK, or more yellow, for instance in the Netherlands and Germany, making breeding for potato flesh colour important. Fig. 1 shows an example of white and yellow flesh colour. The yellow to orange colouring of the potato tuber flesh is caused by the presence of carotenoids (Wolters et al., 2010).

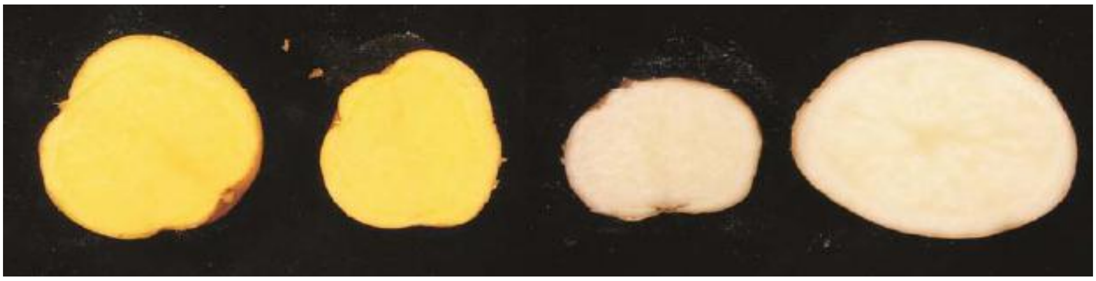


Figure 1: Examples of white and yellow flesh colour in potato

Tuber shape varies from long to compress/round and is measured by the ratio between length and width. Long tubers can be used for French fries while round ones are preferred for crisps. Such traits are controlled by a number of genes and highly influenced by environmental factors . Because cultivated potato is an autotetraploid species, genetic studies are complex, but there are some studies on tetraploid potato as well (D'hoop 2009). However, most genetic studies using biparental crosses are on diploid potatoes (Menendez et al., 2002; Schafer-Pregl et al., 1998). Also in this thesis, we used a diploid backcross population, indicated with 'CxE' where clone C (USW533.7) is a hybrid between *Solanum phureja* and *Solanum tuberosum* and clone E (77.2102.37) is the result of a cross between clone C and *Solanum vernei* (Celis-Gamboa et al., 2002).

The ~omics era: data generation and plant breeding

Researchers in plant breeding are often interested in finding out how certain traits, especially quantitative traits, in plants are regulated in terms of genetic pathways, developmental processes and environmental conditions. These quantitative traits are often difficult to select for, and in some cases it is expensive or difficult to measure the phenotype. With the help of molecular markers it is possible to find statistical associations of genomic regions with the trait variation using analysis of quantitative trait loci (termed QTL analysis). However, there are certain limitations to QTL analysis: we do not directly find candidate genes due to the limited resolution of QTL mapping studies and sometimes we do not find all the genome regions involved, due to the limited power of the QTL studies. In addition, QTL analysis does not show the direct or causal relation to the trait of interest, first of all because we do not necessarily identify the causative gene(s), but also because the influence of the gene is through regulation of certain pathways. These pathways involve proteins, primary and secondary metabolites, and their influence may probably be through many processes interacting with the genes and the genetic and metabolic pathways. In order to shed more light on these other, contributing factors, we also need to study the relationship between the traits of interest and the expression of genes and proteins, and the presence and quantitative variation of metabolites. High-throughput ~omics technologies like microarray (Brazma and Vilo 2000) mass spectrometry (e.g. LC-MS, GC-MS) (Fiehn 2002; Dunn et al., 2005) and protein chips (Aebersold and Mann 2003; Zhu et al., 2003) have gained much interest in the crop sciences. 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 (Joyce and Palsson 2006).

Transcriptomics

Transcriptomics refers to the quantification of mRNA transcripts in a given organism, or in a particular tissue or cell type. Higher abundance of the mRNA transcripts are indicative of higher gene expression of the corresponding gene.

Common technologies for high-throughput analysis of gene expression are: cDNA microarray, oligonucleotide microarray, cDNA-AFLP (Amplified Fragment Length Polymorphism), SAGE (Serial analysis of gene expression) and RNAseq (Ozsolak and Milos 2011; Wang et al., 2009). In this Thesis we use data from cDNA microarrays to study gene expression in combination with phenotypic traits and other ~omics data sets.

Microarray

A cDNA microarray works by using the ability of a given mRNA molecule to bind specifically to, or hybridize to, its original DNA coding sequence in the form of a cDNA template spotted on an array. cDNA microarray experiments typically involve hybridising two mRNA samples, each of which has been converted into cDNA and labelled with its own fluorescent dye (usually a red fluorescent dye, Cyanine 5 (Cy5) and a green-fluorescent dye, Cyanine 3 (Cy3), on a single glass slide that has been spotted with (several thousands of) cDNA probes. Because of competitive binding between the two samples, the ratio of the red and green fluorescence intensities for each spot is indicative of the relative abundance of the corresponding cDNA in the two samples and provides information on the relative expression of the genes.

Metabolomics

Metabolomics is the comprehensive analysis in which metabolites of an organism are identified and quantified (Griffiths and Wang 2009). The components of the metabolome can be viewed as the end products of gene expression that define the biochemical phenotype of a cell or tissue. Currently, some of the technologies available for analyzing a metabolome are: mass spectrometry (MS), nuclear magnetic resonance (NMR), liquid chromatography (LC), gas chromatography (GC), liquid chromatography-mass spectrometry (LC-MS), gas chromatography-mass spectrometry (GC-MS). In this thesis, we used GC-MS and LC-MS data sets which are briefly described below.

Gas Chromatography-Mass Spectrometry (GC-MS)

The GC-MS (Weckwerth 2003) approach is an analytical technique that can be used for plant metabolomics to detect mainly volatiles and primary metabolites such as organic acids and hormones. With GC-MS, first compounds are separated by GC and then transferred online to a mass spectrometer (MS) for further separation and mass detection. The GC can separate metabolites that have almost identical mass spectra (such as isomers), while MS provides fragmentation patterns that differentiate between co-eluting but chemically diverse metabolites. In this way hundreds of different compounds can be detected in parallel, although some can only be detected after deconvolution of overlapping peaks (www.systemsbiology.nl; Griffiths and Wang 2009).

One major limitation of GC is that it can only be used for volatile compounds or compounds that can be chemically transformed into volatile derivatives (derivatisation). The volatile compounds are fractionated by a gas chromatograph, which is usually coupled to a mass spectrometer. The volatile compounds that come off the GC column are not ionized and therefore first need to be ionized for (ion) mass detection and impact ionization, which is an ionization method that actually fragments each gaseous component into ion fragments. Due to the ionization method, each compound results in a specific fragmentation pattern. From the resulting mass spectra, peaks can be quantified based on their mass to charge ratios (m/z values). The fragmentation spectra of individual components are actually used to aid to the deconvolution of overlapping peaks in the total ion trace. In GC-MS each compound is thus characterized by its specific GC retention time and its specific fragmentation pattern (Griffiths and Wang 2009).

In potato breeding, metabolomic studies have progressively increased in importance as many potato tuber traits such as content and quality of starch, chipping quality, flesh colour, taste and glycoalkaloid content have been shown to be linked to a wide range of metabolites (Coffin et al., 1987; Dobson et al., 2008). GC-MS is useful for the rapid and highly sensitive detection of a large fraction of plant metabolites covering the central pathways of primary metabolism (Roessner et al., 2000; Lisec et al., 2006). Untargeted metabolomic approaches by GC-MS have been successfully applied to assess changes in metabolites, evaluate the metabolic response to various genetic modifications (Roessner et al., 2001; Szopa et al., 2001; Davies et al., 2005), and to explore the phytochemical diversity among potato cultivars and landraces (Dobson et al., 2008; 2009).

Liquid Chromatography-Mass Spectrometry (LC-MS)

In Liquid Chromatography-Mass Spectrometry (LC-MS), analytes are separated according to their retention times on an HPLC (high performance liquid chromatography) column, and an MS-method to separate them according to their molecular masses. The LC-MS approach is similar to GC-MS (Weckwerth 2003) except that the samples are not derivatised before analysis, and an HPLC instrument is used to fractionate the samples.

Although the number and range of metabolites that can be quantified with GC-MS is impressive, an even larger number of compounds cannot be determined using GC-MS. For labile compounds, for compounds that are hard to derivatise, or hard to render volatile, LC-MS is more suitable than GC-MS. The main objective for using LC-MS analyses, also the case for this Thesis, is to detect and quantify secondary metabolites responsible for plant secondary metabolism. Due to improvements of LC-MS it is now possible to measure isoprenes, alkaloids, phenylpropanoids, glucosinolates, flavonoids, saponins and oxylipins. However, optimal extraction procedures differ for each class of compounds, and it is unlikely that a single extraction allowing accurate quantification for all the above compounds will be achievable.

Proteomics

Proteomics is the comprehensive, quantitative description of protein expression or protein-level measurements and its changes under the influence of biological (such as disease) or environmental conditions. There are various methods for generating proteomics data such as using gel electrophoresis (for example: 1-Dimensional, 2-Dimensional gel electrophoresis), or using MS based proteomics such as LC-MS, MS-MS. In this thesis, we used 2D difference gel electrophoresis (DIGE) for quantification of the protein expression.

2D difference gel electrophoresis (DIGE)

Plant tissue samples are labeled prior to electrophoresis with fluorescent dyes Cy2, Cy3 or Cy5. Two samples are then mixed prior to IEF (isoelectric focusing) and resolved on the same 2D gel (Viswanathan et al., 2006). The two dyes allow a comparison of the two samples on the same gel directly. Instead of comparing two experimental samples, each sample in a larger experiment can also be compared to a reference sample containing a mixture of all samples in the experiment. This pooled standard can then be used to normalize protein abundance measurements across multiple gels in an experiment. As a consequence, each gel will contain an image with a highly similar spot pattern, simplifying and improving the confidence of inter-gel spot matching and quantification.

So far, there have been only limited potato proteomics studies, but the few that have been performed show that proteomics research might add to the identification of genes involved in certain processes (Urbany et al., 2012).

Genetical genomics

Allelic variations in the DNA sequence of genes and their regulatory regions underlie most of the phenotypic variation that has been exploited in modern crops and specifically in plant breeding (Bryan et al., 2000; Masouleh et al., 2009). Quantitative Trait Loci (QTL) (Paterson et al., 1988; Collard et al., 2005; Kearsey 1998; Kearsey and Farquhar 1998; Mackay 2001) mapping allows the identification of such genomic regions involved in quantitative traits in a segregating population (F1 of a cross between two heterozygous plants of a cross pollinator, F2, backcross, doubled haploids or recombinant inbred lines). We can consider quantified gene expression levels as molecular quantitative traits that can also be used in a QTL analysis. This approach is called ‘genetical genomics’ (Jansen and Nap 2001). More recently, metabolite and protein abundances have also been considered as molecular quantitative traits and analyzed by performing QTL analyses (Doerge 2002; Schadt et al., 2003). QTLs of gene expression profiles are denoted as expression QTLs (eQTL) (Schadt et al., 2003). By analogy, QTLs for proteomics data are called protein QTLs (pQTLs) and QTLs for metabolomics data metabolite QTLs (mQTLs) (Keurentjes et al., 2006; Kliebenstein 2007; Acharjee et al., 2011). The objective of a genetical genomics study is to identify genomic regions associated with observed variation in and between molecular quantitative traits such as gene expression. This approach can be used to identify regulatory genes that regulate other sets of genes as well as to derive information on genetic pathways and the relationship between regulatory and functional genes. Although the genetical genomics approach has successfully been applied to understand the genetic basis of ~omics data (Keurentjes et al., 2006) there are still limitations: the mapping resolution is often not high enough to identify causal genes underlying detected QTLs (phQTLs, eQTLs, mQTLs or pQTLs) directly. However from using the different ~omics platforms, additional information will be available which can aid in identifying functional relationships, regulatory functions, DNA sequence positions and pathways.

Data integration and networks with ~omics

The rapid advances in '~omics' technologies (genomics, transcriptomics, proteomics, and metabolomics) provide an opportunity to better understand the organization principles of cellular functions at different levels such as metabolite, gene or proteomic levels, but we have to link quantities of the metabolites, proteins and gene expression to a phenotype of interest. Therefore an integrative approach is needed for this (Fukushima et al., 2009). We can consider different approaches:

- Linking a phenotypic trait to a single ~omics data set (whether transcriptomics, proteomics, LC-MS or GC-MS)
- Linking a phenotypic trait to multiple ~omics data sets simultaneously
- Linking two (or more) ~omics data sets to one another

Linking a phenotype to an ~omics data set

High-throughput technologies such as microarrays, LC-MS, GC-MS are commonly used to study the behaviour of genes, proteins and metabolites and typically generate large data sets. Here 'large' is referring to the number of genes, proteins, metabolite peaks compared to the much smaller number of samples (individuals) being tested in any given study and hence it creates a high-dimensional, multivariate data set.

In plant breeding research, phenotypic data of interest could be disease scores, growth characteristics, agronomical traits or quality traits such as flesh colour of potato, phosphate content etc. Such phenotypic data can be scored on a continuous scale, an ordinal scale or as binary scores. To link such phenotypes with ~omics data one could think of a regression approach (in case of a binary response, one could use a classification or a logistic regression approach) where the phenotypic trait is considered as a response variable and the ~omics data set as a predictor set. This makes sense as we are usually interested in the phenotypic trait as predicted from the molecular profile variation between individuals in a population. However, in such data sets the number of variables (p) is larger than the number of individuals (n) and there will be collinearity due to p>>n (Kiers and Smilde 2007) but also because of high correlations among sets of variables due to common biological functions. Because of this, we cannot invert the variance-covariance matrix to estimate regression coefficients and hence traditional statistical methods such as multiple linear regression cannot be applied.

Therefore we need other methods that can be used despite the overabundance of candidate variables associated with a response variable. Traditional methods such as forward selection or forward stepwise regression could be applied but they have some limitations. In forward selection, one starts with an empty model and adds variables to the model, each time the one that gives the best improvement to the model, given the variables already present. When the improvement is no longer statistically significant, this process is stopped so that a subset of the explanatory variables is included in the model. In stepwise regression, in each step it is evaluated whether removing or adding a variable gives the best improvement of the model (Hastie et al., 2001).

The mean square error (MSE) of a regression model can be decomposed into two components: the square of the bias (difference between the estimate and the expectation of a parameter) and the variance of the parameter estimate. When there are many correlated variables in a linear regression model, their regression coefficients can become poorly determined and their estimates will exhibit high variance. A large positive regression coefficient of one variable can be canceled out by a similarly large negative coefficient of its correlated cousin. The recent statistical literature shows that approaches using regularization or penalization or shrinkage (Ghosh 2008; Hastie et al., 2001) are the most preferred in this context.

Regularization methods impose a penalty on the size of the regression coefficients; by doing so, the variance-covariance matrix can be inverted and hence we can estimate regression coefficients; those estimated coefficients will be shrunken towards zero or exactly zero due to the imposed size constraint (penalty).

Preprocessing and standardization of ~omics data sets

Preprocessing of ~omics data sets is done mainly to remove known systematic errors (other than the treatments applied or the variation we want to investigate), to improve comparability among samples. For example, background correction in a transcriptomics study is meant to remove signal that is always there, regardless of the level of gene expression. In metabolomics (LC-MS or GC-MS) data sets preprocessing steps could include baseline correction, alignment of peaks, peak detection etc. Before a statistical analysis of ~omics data, the data set is usually ^{10}log or ^{2}log transformed. The motivation for the log transformation is that the distribution of expression level is typically asymmetric with a long tail at the high expression end. Many statistical tests require variables to follow a Gaussian (normal) distribution. Other advantages are that it might reduce heteroscedasticity (i.e. increasing variance with increasing mean values) and improve linearity of effects (Van den Berg et al., 2006). Depending on the statistical methods, further preprocessing may be required. One of the ways to do this, is standardization of the variables, also called autoscaling. Autoscaled variables have a mean of zero and a variance (and also standard deviation) of one, thereby if weights are induced by differences in variance or mean, this effect is taken away.

Prediction

The availability of high-throughput ~omics (transcriptomics, proteomics and metabolomics) and molecular marker data makes it possible to infer and predict direct relationships between a quantitative trait and an ~omics data set. For prediction often regression methods are used. To assess the goodness-of-fit of a regression model, R^2 can be used; however, R^2 does not quantify the prediction performance for *new* data. Therefore we are interested in assessing the quality of the model in terms of prediction; resampling methods (for example: cross-validation) can be used in order to quantify the accuracy of prediction of so-called new data (data not used to fit the model) from the actual data (data in hand used to fit the model).

In this thesis we used cross validation since an independent test set was not available. In such situations, we can apply k-fold cross validation by dividing the data set randomly into a number of folds that, together, make up the whole data set. For example, in ten-fold cross-validation (where k=10) the data set is divided into ten folds: nine tenth is then used for fitting the regression model (called the “training set”), and one tenth portion is left out and used to test the predictions (called the “test set”). All tenth parts are rotated so that they each have been the test set exactly once, so that after ten rounds every individual sample has been in the test set exactly once. In addition to this, many different random divisions into ten subsets can be generated to repeat the ten-fold cross validation, thereby avoiding that the prediction is just based on one particular division of the data set.

To evaluate the performance of regression models (Mevik et al., 2004) two parameters are used: Goodness of fit (R^2) and the mean squared error of prediction (MSEP) (Mevik et al., 2004). Goodness of fit (R^2) is used to describe how well the predictions fit a set of observations. It is a measure for the proportion of variability in a data set that is accounted for by the statistical model. The usual R^2 from a linear regression is just a measure of goodness-of-fit of the data at hand (training data), but is not usually valid for future predictions (test data). However, in the cross validation, we can estimate the R^2 on the test set, as this set is used as an independent data set.

The mean squared error of prediction (MSEP) is obtained by averaging the squared prediction errors (differences between observed and predicted values) of the test samples. A lower MSEP value corresponds to a better predictive model.

Ranking of variables or variable selection

In high-dimensional ~omics data sets, we may be interested to find a relevant smaller subset of variables which, combined, are associated with the response (a phenotypic trait, usually). Procedures to find such smaller subsets are called *variable selection* procedures. By doing this, it is possible to reduce the dimensionality of the data set and perhaps also to get rid of some or even many noise variables (variables which have no predictive power for the response variable) in the data set. Some of the regularization regression methods which select a single or a smaller subset of variables from a set of correlated variables are adaptive LASSO (Zou 2006) and Bayesian LASSO (Park and Casella 2008). Some of the examples of regularized regression methods which also do subset selection but select also groups of correlated variables are: LASSO (Least Absolute Sum of Squares Operator) (Tibshirani 1996), elastic net (Zou and Hastie 2005), sparse PLS regression (SPLS) (Chun and Keles 2009), group lasso (Yuan and Lin 2004) and a sparse group lasso (Friedman et al., 2010).

There are also methods which do not perform variable selection, but instead use all variables in the prediction. These methods include ridge regression (Hoerl and Kennard 1970), principal component regression (PCR) (Massy 1965), partial least squares (PLS) regression (Wold 1975), and Random Forest (RF) regression (Breiman 2001).

These methods still allow ranking of the variables based on the size of the regression coefficients or other measures of variable importance such as the Gini index in Random Forest (Breiman 2001). However, they differ in the criterion and/or the function of the regression coefficients that is being penalized. For example, in lasso it is the sum of the absolute values of the regression coefficients, in ridge regression the sum of the squares of the regression coefficients, and in elastic net a combination of both. In PLS (partial least squares) and PCR (principal component regression) the reduced number of principal/PLS components imply a penalization. The regression coefficients are shrunken due to this penalization (Hastie et al., 2001).

Some of the regularization methods have only a single parameter that needs to be optimized, for example: partial least squares regression, lasso, principal components regression, ridge regression. In other methods there might be two parameters, for example the two penalty parameters in case of elastic net regression; Random Forest, and sparse partial least squares also have two parameters that need to be optimized.

Linking a phenotypic trait to multiple ~omics data sets

If we have multiple ~omics data such as transcriptomics and metabolomics (LC-MS or GC-MS); transcriptomics and proteomics; metabolomics and proteomics; or transcriptomics, metabolomics and proteomics data sets and a trait of interest, we are in a different situation than just linking each of them separately to the trait (explained in the previous section).The aim is to find relationships among these multiple different levels of regulation with a connection to a phenotype of interest, for example to relate phenotype (e.g. potato tuber flesh colour) to gene expression *and* metabolite levels. Here we can consider two possibilities: fuse multiple ~omics data sets into a single data set and treat the variables as if they are from a single data matrix and regress the phenotype on the variables in the data matrix. The other one is to treat the ~omics data sets separately and regress a phenotype separately on each one, select subsets of variables for each and then evaluate the relationships among the variables selected from the different data sets.

Linking two or more ~omics data sets

When comparing multiple ~omics data sets, we are not dealing with a single response variable versus a multivariate ~omics data set, but rather two or more multivariate data sets for which we want to find relationships. In this situation we are also not predominantly interested in predicting one data set from another one (or multiple others), but we consider the relationships between the data sets in any direction (e.g. we want to find relationships between a transcriptomics data set with a proteomics data set).

Integrating two data sets from different analytical platforms can enable an improved understanding of some underlying biological mechanisms and interactions between different functional levels. The advantage of this approach is that we can get integrated results, e.g. from multiple correlated metabolites and genes together at the same time**.** Different attempts have been made to integrate multiple ~omics data sets from different species such as metabolomics and proteomics in *Arabidopsis thaliana* using principal components analysis (PCA) and independent components analysis

(ICA) (Wienkoop et al., 2008), and transcriptomics and metabolomics in *Arabidopsis thaliana* using orthogonal partial least squares regression (O2PLS) (Bylesjo et al., 2007), and transcriptomics, metabolomics and proteomics in grapevine berry also using O2PLS (Zamboni et al., 2010). Other related statistical methods can be canonical correlation analysis (CCA) which is an exploratory statistical method to highlight correlations between two data sets acquired on the same experimental units (n) but in case of ~omics data sets, where the number of variables (p) is much larger than the number of experimental units (n), a regularized version of canonical correlation analysis (CCA) should be applied to overcome p>>n issues (González et al., 2008). A regularized CCA allowing variable selection, sparse canonical correlation analysis was developed and applied by Waaijenborg and Zwinderman 2009 to a gene-expression microarray and a data set of DNA-markers.

A two-way variant of partial least squares (PLS) regression, PLS2 (Wold 1966) is capable of handling two or more multivariate data sets for which we want to find relationships. PLS2 has been successfully applied to biological data, such as gene expression, integration of gene expression and clinical data (with bridge PLS, Gidskehaug et al., 2007). A type of PLS2 called sparse PLS2 (Lê Cao et al., 2008, 2009) was developed to simultaneously integrate and select variables using lasso penalization (Tibshirani,1996). This method was applied to two different transcriptomics platforms: cDNA and Affymetrix chips where gene expression was measured on sixty cancer cell lines with both platforms. Such methods (CCA, O2PLS, PLS2) are relevant in this thesis as one could find multiple different levels of regulation and relationships across ~omics data sets, for example: gene expression vs. proteomics data sets, gene expression vs. metabolomics (LC-MS or GC-MS) data sets.

Networks

Genes and their gene products interact with one another and with other molecules (proteins, metabolites etc.) in a regulatory web of cause and effect including different feedback loops. Gene induction or repression occurs through the action of specific proteins, which are in turn products of certain genes, but gene expression can also be affected directly by metabolites or proteins, or through gene-gene, gene-metabolite, gene-protein, metabolite-protein or gene-protein-metabolite complexes and their interactions. Such interactions can be modeled in a cellular network (Barabási et al., 2004 and 2011; Bernardo et al., 2005; Bansal et al., 2006) where genes, metabolites and proteins can be represented as nodes in a network and the strength of their interactions as edges in the network. Such networks come in two types: inter-level and intra-level. In the intra-level category, we focus on relationships within one particular molecular domain; for example, a network consisting only of genes is called a gene regulatory or gene co-expression network; networks containing only metabolites are denoted as metabolite networks (Lacroix et al., 2008; Terzer et al., 2009; Han 2008; Fiehn et al., 2003). Inter-level networks consist of multiple types of molecules such as expression of genes and metabolites (Nikiforova et al., 2005; Acharjee et al., 2011), gene expression and proteins (Russo et al., 2010), metabolites and proteins (Yamada et al., 2009), expression of genes, metabolites and proteins (Yuan et al., 2008), genes, metabolites, proteins and phenotypes. In Fig. 2, we show an approach of inter-level networks by integrating ~omics data with a phenotype of interest and associations among ~omics data sets. To build an inter or an intra-level network different methods can be applied: Pearson correlation, mutual information, partial correlations etc. A detailed review about the methods can be found in Markowetz and Spang 2007. In this thesis, we used Pearson correlation coefficients (in chapter 3) for building a network with genes, metabolites and a phenotypic trait.

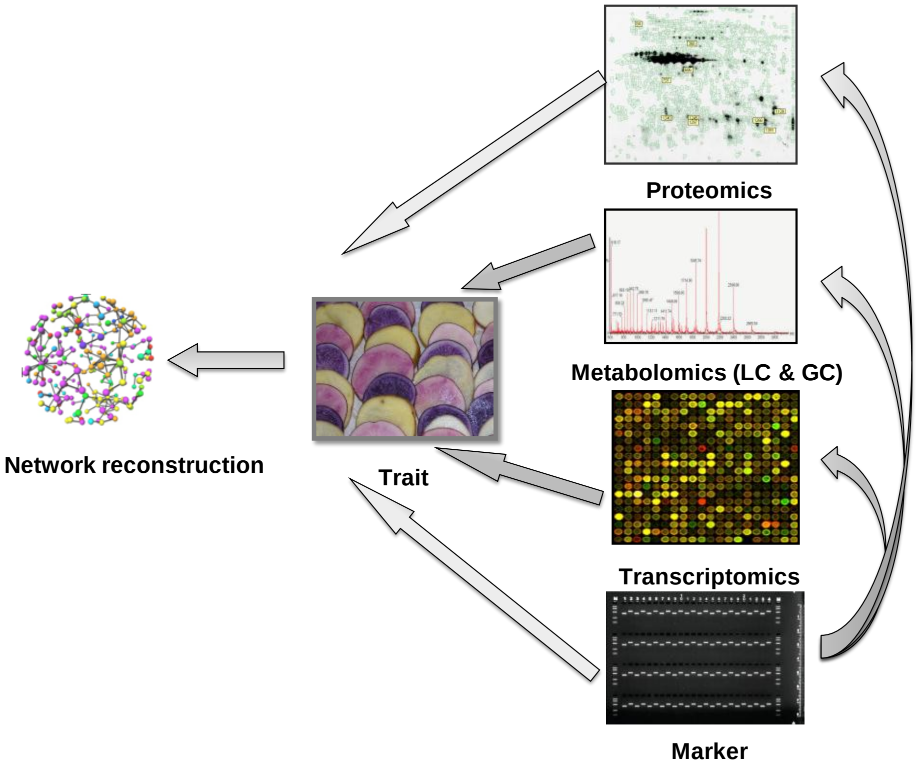


Figure 2: Integration of ~omics data sets (transcriptomics, metabolomics and proteomics) and markers with one trait (here, flesh colour) at a time to identify important metabolites, genes and proteins associated with flesh colour, followed by a correlation network analysis with a trait, genes, proteins and metabolites.

Objectives and outline of this thesis

In this thesis we used a segregating diploid potato population (CxE) that has also been used for mapping and QTL analysis. Over the years and different research projects, much data has been accumulated: molecular marker data, phenotypic data (e.g. developmental traits, tuber quality traits), microarray data, metabolomics data (LC-MS and GC-MS) and 2D DIGE proteomics data. The current challenge and the subject of this thesis is to extract biologically meaningful associations from these data sets and relate these to phenotypes of interest, to study the methodology used to find these associations and to obtain an idea about the general reliability of the used methods and associations.

The main objective of this thesis is to link a particular phenotype to ~omics data to finally end up with a minimum set of markers (irrespective of being metabolite, transcript or protein) which can predict a quantitative trait. This objective can be subdivided into the following parts: to link phenotype to omics data, ~omics to ~omics data sets and the study of the genetics of (quantitative) traits and ~omics data sets.

A further goal is to compare statistical methods suitable for ~omics data sets in terms of the prediction, correlations among variables, and ranking of genes, metabolites or proteins based for example on regression coefficients or Gini index in predictive models. To achieve this goal we considered quality traits in potato (such as tuber flesh colour, enzymatic discoloration, phosphate content, cold sweetening traits, tuber shape and starch gelatinization). We analyzed different ~omics data sets: transcriptomics, metabolomics (LC-MS & GC-MS) and proteomics data sets in the C x E potato population. These results are described in the seven chapters of this thesis.

In chapter 2 we study different regression methods to link a quantitative phenotypic trait to a metabolomics data set (the predictor data set). These regression methods are all methods that can be used in typical ~omics situations with large numbers of variables and smaller numbers of samples. We compare the methods in terms of mean square error of prediction, goodness of fit, variable selection and the ranking of the variables.

In chapter 3, we study potato tuber quality traits in relation to transcriptomics and metabolomics (LC-MS) data sets, using a Random Forest approach, and we select a subset of metabolites and transcripts that show an association with the quality traits. We construct a Pearson correlation network for two of the quality traits, flesh colour and enzymatic discoloration, with gene expression data and metabolites, leading to the integration of known and uncharacterized metabolites with genes already known to be associated with the carotenoid biosynthesis pathway. We show that this approach enables the construction of meaningful networks with regard to metabolite pathways.

In chapter 4, we use GC-TOF-MS data sets to identify genetic factors underlying variation in primary metabolism in a mapping population. We perform a QTL analysis for starch and cold sweetening related traits and infer links between these phenotypic traits and primary metabolites. We apply Random Forest regression to find significant associations between phenotypic and metabolic traits. We confirm putative predictors in an independent collection of potato cultivars. Our results show the value of combining biochemical profiling with genetic information to identify associations between metabolites and phenotypes. This approach reveals previously unknown links between phenotypic traits and primary metabolism.

In chapter 5, we perform a proteomics analysis of potato tubers in order to obtain an insight into the relationships between protein traits and tuber quality traits such as enzymatic discoloration, starch and cold sweetening related traits. We use genetic information through QTL co-localizations and a Pearson correlation study between protein traits and quality traits. We show hot spot areas for protein QTLs consistent between data sets of two growing years (2002 and 2003). We report the first attempt for identification of protein spots of which the QTLs co-localize with quality traits.

In chapter 6, we perform an integrated analysis over all the ~omics data sets. Here we study the relationship between phenotypic traits (tuber quality traits) and multiple related ~omics data sets simultaneously. We apply a genetical genomics approach to find regions of the genome explaining quantitative variation in the transcripts, metabolites and proteins predictive for quality traits. We present an approach to find a limited set of genes, metabolites and proteins for which the association to the trait is a functional relationship. First, we select subsets of genes, metabolites (LC-MS and GC-MS) and proteins showing a significant association with phenotypic traits, using Random Forest. Then variation in the expression of selected genes or in concentration of metabolites and proteins is mapped as eQTLs, mQTLs and pQTLs across the genome. Per trait, genomic regions associated with the trait are identified; in a third step, representatives of genes, metabolites and/or proteins are selected for an integrated network analysis. This integrated analysis results in a list of a minimum set of candidate genes and underlying metabolic pathways possibly linking genes, metabolites and proteins in different genomic regions underlying trait variation.

Chapter 7 provides a general discussion on all the findings. Suggestions for future approaches for the dissection of traits and ~omics data sets are given. We discuss appropriate statistical methodologies for ~omics studies in terms of prediction and variable selection. We also discuss how plant breeders can use the integrated knowledge about marker and ~omics data for selecting a trait.

Chapter 2

Comparison of regularized regression methods for metabolomics data

Animesh Acharjee[1,2], Richard Finkers[2,3], Richard GF Visser[2,3] and Chris Maliepaard[2,3]

[1]Graduate School Experimental Plant Sciences, Droevendaalsesteeg 1, 6708 PB Wageningen, The Netherlands

[2] Wageningen UR Plant Breeding, Wageningen University and Research Center, P.O. Box 386, 6700 AJ Wageningen

[3]Centre for BioSystems Genomics, PO Box 98, 6700 AB, Wageningen, The Netherlands

To be submitted

Abstract

In this study, we compare methods that can be used to relate a phenotypic trait of interest to an ~omics data set, where the number or variables outnumbers by far the number of samples. We apply univariate regression and different regularized multiple regression methods: ridge regression (RR), LASSO, elastic net (EN), principal components regression (PCR), partial least squares regression (PLS), sparse partial least squares regression (SPLS), support vector regression (SVR) and Random Forest regression (RF). These regression methods were applied to a data set from a potato mapping population where we predict potato flesh colour from a metabolomics data set. We compare the methods in terms of the mean square error of prediction of the trait, goodness of fit of the models, and the selection and ranking of the metabolites. In terms of the prediction error, elastic net performed better than the other methods. Different numbers of variables are selected by the methods that allow variable selection but seven variables were in common between LASSO, EN and SPLS. SPLS performed better than EN with respect to the selection of grouped correlated variables. We developed a web application that can perform all the described methods and that includes a double cross validation for optimization of the methods and for proper estimation of the prediction error.

Key words: metabolomics, regularized regression, variable selection, variable ranking.

1 Introduction

High-throughput technologies like microarray (Brazma and Vilo 2000; Gaasterland and Bekiranov 2000) mass spectrometry (*e.g*., LC-MS, GC-MS) (Fiehn 2000; Dunn et al., 2005) and protein chips (Aebersold and Mann 2003; Patterson and Aebersold 2003; Zhu et al*.,* 2003) have gained much interest in the biological domain. These techniques allow one to measure thousands of variables (genes, metabolites, proteins) simultaneously. The data generated by these techniques are often denoted as ~omics data (Joyce and Palsson 2006). These data sets are generally very large in terms of the number of variables (p) and often small in terms of the number of the biological samples (n). In statistics, this problem is often termed as the *"large p and small n problem" (p>>n).* In such wide data sets there will be collinearity due to p>>n (Kiers and Smilde 2001) but also because of high correlations due to common biological functions (*e.g*. metabolites in the same pathway).

In many of these ~omics situations one wants to find functional relationships between a phenotypic trait of interest and the ~omics variables and often the interest would also be in selecting a smaller subset of the variables that have good prediction of the trait.

In traditional statistical methods, multiple linear regression techniques are used for prediction situations such as outlined above, but, due to the high collinearity, these methods cannot be applied. Therefore we need different approaches: penalization regression methods or machine learning methods.

We wanted to compare the different methods on real data, but we still wanted to be able to infer whether results were biologically meaningful, so we chose a trait for which a fair amount of information is already available, including a possible relationship to underlying metabolites. Therefore, we considered potato tuber flesh colour as the phenotypic trait of interest, the response in our regression, and a large metabolomics data set as the set of predictor variables. We apply a double cross validation scheme to include optimization of any hyperparameters needed in the models, and allow estimation of prediction error.

We apply different regression methods: ridge regression (RR) (Hoerl and Kennard 1970), LASSO (Tibshirani 1996), elastic net (EN) (Zou and Hastie 2005), principal component regression (PCR) (Massy 1965), partial least squares regression (PLS) (Wold 1975) sparse PLS regression (SPLS) (Chun and Keles 2009), support vector regression (SVR) (Vapnik 1995) and Random Forest regression (RF) (Breiman 2001).

We use univariate regression as a reference and compare the results of univariate regressions with multiple regression methods. We also study the properties of these multivariate methods both from a theoretical point of view as well as their performance in practical situations in terms of the variable selection (Saeys et al., 2007) or ranking of the variables, grouping of correlated variables in variable selection (Zou and Hastie 2005), and the prediction error. Regarding the grouping of correlated variables, we are interested in finding out whether in the variable selection methods (LASSO, EN, SPLS) variables are selected as a group or not, and we also compare regression coefficients of these correlated variables.

So far in literature, comparison studies are usually focused on classification methods instead of regression methods (Hendriks et al., 2007) and data used in these studies often were transcriptomics data (Bøvelstad et al., 2007, 2009). In the context of regression, Kiers and Smilde 2001 did a comparison of various multiple regression methods on simulated data with collinear variables but their study was mainly focused on prediction and comparison of the regression coefficients when predictor variables are collinear. Menendez et al., 2012, reported comparison of stepwise linear regression, LASSO, EN and RR but did not cover other penalization methods such as SPLS, PLS, PCR, RF and SVM. We compare these methods (RR, EN, LASSO, SPLS, RF, SVM, PLS, PCR) in terms of mean square error of prediction, goodness of fit, variable selection and the ranking of the variables. In addition, we developed a web application with all the methods mentioned including the double cross validation procedure. This website can be accessed from : http://www.plantbreeding.wur.nl/omicsFusion/

2 Materials and methods

2.1 Plant material

Ninety-one individuals from a diploid mapping population of potato were used in this study. Clone C is a hybrid between *Solanum phureja* and *Solanum tuberosum*. Clone E is the result of a cross between Clone C and *Solanum vernei* (Celis Gamboa et al., 2003). All clones were grown in the field, Wageningen, The Netherlands in 1998. For each genotype, tubers from two plants were collected and representative samples from these tubers, of each genotype, were used for phenotypic analysis directly after harvest, and for LC-MS.

2.2 Evaluation of phenotypic traits

Many quality traits were collected for this potato population (Celis-Gamboa 2002; Celis-Gamboa et al., 2003; Werij et al., 2007). In this study we used one well-studied phenotypic trait (potato tuber flesh colour) allowing better to compare methodology and to be able to verify the results that we found. Potato tuber flesh colour was visually scored on a scale from 1 (white) to 9 (dark yellow/orange) in three repeats consisting of two plants each. Flesh colour scores were then averaged over the three repeats.

2.3 Data preprocessing

For metabolomics analysis the exact same material (potato tubers of the same genotypes) was used for Liquid chromatography–time of flight mass spectrometry (LC-QTOF MS) analysis which resulted in over 16,000 individual mass peaks. Mass peak signals below background were removed resulting in about 10,000 remaining mass peaks. The next step was to make a selection of these 10,000 peaks based on skewness of the data and all mass peaks with a skewness score below -2 and above 2 for the progeny and a score below -1 and above 1 for the parental repeats were discarded. The signal intensities of the 1,100 remaining mass peaks were then correlated to the available quality trait data of this population, in order to obtain the most interesting metabolites *i.e.* the metabolites linked to quality traits, P-values of these correlations were calculated using Student's t-test. A number of 163 mass peaks with the highest significance ($p<0.0005$) were selected. Before analysis the metabolite data was ^{10}log transformed for symmetry and then autoscaled. Autoscaled variables have a mean of zero and a variance (and also standard deviation) of one, thereby giving all variables (mass scan numbers) an equal weight in the analysis. LC-MS peaks are characterized by their mass and scan number (mass_scan).

2.4 Statistical methods for regression in p >> n situations

2.4.1 Methods used

We compared the prediction, variable selection and ranking of variables. In this section, we first review the regression model in these eight methods. For all methods, values for one or more tuning parameters needed to be chosen. This was done using tenfold cross-validation, described in section of criteria for comparison of the methods.

2.4.2 Regression Methods:

Regression methods are essentially curve-fitting approaches. When there is one response variable and one predictor variable, simple linear regression consists of finding the best straight line relating the response to the predictor variable. In case of multiple predictors, a hyperplane is fitted. The usual criterion, the least squares criterion, minimizes the sum of squared distances between the observed responses and the fitted responses from the regression model (Montgomery and Peck 1991). We can represent the least squares criterion as:

$$\min_{\beta} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

Where; y= response vector (here: flesh colour); β=regression coefficients; x=predictor variables (the log intensity values of the mass_scans from LC-MS data measured over different samples)

Here we are describing nine different regression methods which were applied to relate potato tuber flesh colour to the LC-MS data set

2.4.3 Univariate regression

Univariate regression was used as a reference. We compare the variable selection and ranking of variables in the multivariate regression methods to the results from the univariate regressions of flesh colour to each of the individual LC-MS peaks. Univariate regression with a FDR (False discovery rate) adjustment was done according to the procedure of Benjamini and Hochberg 1995.

2.4.4 Penalization or shrinkage methods

Shrinkage methods, also called penalization methods, impose a penalty on the size of the regression coefficients. The penalty term is also called a ‘regularization parameter’. We have grouped the methods according to the type of penalty applied to the regression coefficients. The mean square error (MSE) of a regression model can be decomposed into two components: the square of the bias (difference between the estimate and the expectation of a parameter) and the variance. In situations with high collinearity (p>>n), regression models usually have a very large variance and the MSE will mainly be determined by this large variance. Therefore, in such situations it can be advantageous (lower MSE) to accept some bias if it is allows us to decrease the variance by considerable amount (Hastie et al., 2001). Penalization methods impose a bias by applying a penalty to the regression coefficients.

2.4.5 Continuous penalization methods

In this category of regression methods, shrinkage factors can take any value between zero and infinity. LASSO, RR and EN belong to this category. The value of the shrinkage parameter decides the amount of penalty applied to the regression coefficients. We use tenfold cross validation (Hendriks et al., 2007) to choose the optimum penalty value; this will be discussed in detail in the section about criteria for comparison of the methods

2.4.6 Ridge regression (RR)

Ridge regression (Hoerl and Kennard 1970) shrinks the regression coefficients by imposing a penalty on the sum of squares (L2 norm) of the regression coefficients.

$$\min_{\beta} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda_2 \sum_{j=1}^{p} \beta_j^2$$

The left part of the term shown above is the usual least squares criterion. In the right part, λ_2 is a shrinkage factor applied to the sum of the squared values of the regression coefficients. The larger the value of λ_2 , the heavier the penalty on the regression coefficients and the more they are shrunk towards zero. In ridge regression all the regressor variables stay in the model since regression coefficients do not become exactly zero (that would be equivalent to variables dropping out of the regression model). Ridge gives equal weight to absolutely correlated variables in the data set (Hastie et al*.,* 2001).

2.4.7 LASSO

The LASSO (Least Absolute Shrinkage and Selection Operator, Tibshirani, 1996) is another regularization method but here the penalty is applied to the sum of the absolute values of the regression coefficients, the L1 norm. Mathematically, we can write this in the following way:

$$\min_{\beta} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda_1 \sum_{j=1}^{p} \left| \beta_j \right|$$

Again, the left part of the term is the normal least squares criterion. The right part now is the penalized sum of the absolute values of the regression coefficients. Similar to ridge regression, the shrinkage parameter (λ_1) has to be decided on and again we use tenfold cross validation for this. Penalizing the absolute values of the regression coefficients has the effect that a number of the estimated coefficients will become exactly zero which means that some regressors drop out of the regression model so that a LASSO fitted model can consist of fewer variables than the number of available regressors. In other words, LASSO can implicitly perform variable selection. The number of selected variables is upper limited by the numbers of samples (n). In case of absolutely correlated variables LASSO just selects one and ignores the rest in the group (Hastie et al., 2001).

2.4.8 Elastic net (EN)

Elastic net (Zou and Hastie 2005) is a combination of LASSO and ridge regression. It uses both a ridge penalty (penalty on the sum of the squares of the regression coefficients) and a LASSO penalty (on the sum of the absolute values of the regression coefficients):

$$\min_{\beta} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda_1 \sum_{j=1}^{p} \left| \beta_j \right| + \lambda_2 \sum_{j=1}^{p} \beta_j^2$$

In elastic net, we optimize both penalty parameters simultaneously using tenfold cross validation. Variable selection is encouraged by the LASSO penalty (λ_1) and groups of correlated variables get similar regression coefficients. Groups of correlated variables are either in or out of the model (Zou and Hastie 2005). In contrast with LASSO, the number of selected variables is not limited by the number of individuals.

2.4.9 Discrete penalization methods:

Partial least squares (PLS) and Principal component regression (PCR) are based on latent variables or components which are linear combination of the original variables. For both methods it is essential to select the optimum numbers of latent components for prediction of the response variable. We used tenfold cross validation to choose the optimum number of latent components based on the smallest mean square error of prediction (MSEP) value. The number of latent components can only take discrete values, hence these methods are discrete penalization methods.

2.4.10 Principal component regression (PCR)

Principal component regression (Jolliffe 1982) is a combination of principal components analysis (PCA) and multiple linear regression. First, PCA is done on all original regressors and each component (latent variable) is represented by a linear combination of the original variables. The number of latent variables (components) is chosen by tenfold cross validation and the response is regressed on the selected latent variables. These latent variables in PCA are uncorrelated. In PCR the principal components are found by maximization of the variance in the predictors; the covariance of the predictors with the response variable is not taken into account.

2.4.11 Partial least squares (PLS)

Partial least squares (PLS) (Wold 1975; Geladi and Kowlaski 1986; Hoskuldson 1988) is a method to relate a single response variable or a matrix of response variables to a matrix of regressor variables. Here, we are considering only a single trait as the response. PLS is a dimension reduction method like PCA, but it uses a different criterion: maximization of the covariance between the latent variables and the response. As a consequence, usually fewer components are required for prediction as compared to PCR. The optimum number of latent components is chosen by tenfold cross validation. Since the optimum number of latent components is a discrete number, this method is also a discrete penalization method. Like in PCR, latent variables in PLS are also uncorrelated.

2.4.12 Hybrid penalization method

In this section, we consider a method in which two different types of penalties (continuous and discrete) are applied simultaneously.

2.4.13 Sparse partial least square (SPLS)

SPLS (Chun and Keles 2009) is a combination of two different penalties. The continuous penalty is a LASSO penalty and discrete penalization is achieved by PLS. Variable selection is achieved by LASSO, dimension reduction by PLS. The respective hyperparameters *i.e.* the number of PLS components and the size of the LASSO penalty are optimized simultaneously by tenfold cross validation. As in normal PLS, each of the latent components is a linear combination of the original variables.

2.4.14 Machine learning methods

The goal of machine learning is to build a computer system that can adapt and learn from experience (Dietterich 1999). Machine learning methods can handle data which are not normally distributed whereas the methods mentioned above assume normality. Machine learning methods can also handle nonlinear relationships between response and predictor variables.

2.4.15 Support vector machine (SVM)

The support vector machine (SVM) (Vapnik 1995) was originally developed in a classification (Demiriz et al*.,* 2001) context and maximizes predictive accuracy while avoiding overfitting(Hastie et al., 2001; Cristianini and Shawe-Taylor 2000) to the data.Two parameters such as epsilon (insensitive zone) and regularization parameter "C" are optimized (Vapnik 1995). However, the methodology can also be used in a regression model (Cristianini and Shawe-Taylor 2000). Mathematically, given the input data $\{(\underline{x}_1, y_1), \ldots., (\underline{x}_n, y_n)\}$, we want to find a function which will fit the following equation:

$$f(x) = wx + b,$$

Where w is a weight vector and b is a constant

The goal of support vector regression (SVR) is to find a function f(x) that has at most ε deviation (Cristianini and Shawe-Taylor 2000) from the actually obtained targets (response) for all the predictors, and at the same time minimizes the distance between predicted and target values. SVR does not encourage grouping or variable selection.

2.4.16 Random Forest (RF)

A Random Forest (Breiman 2001) is a collection of unpruned decision trees (Hastie et al*.,* 2001), usually developed for a classification purpose but this method can be applied in a regression context as well (Segal 2004). A RF model is typically made up of hundreds of decision trees. Each decision tree is built from bootstrap samples of the data set. That is, some samples will be included more than once in the bootstrap sample, and others will not appear at all. Generally, about two thirds of the samples will be included in this training dataset, and one third will be left out (called the out-of-bag samples or OOB samples). In RF regression the prediction error is calculated as the average prediction error over OOB predictions. Variable importance (Breiman 2001) can be quantified in RF regression. Variables used which decrease the prediction error obtain a higher variable importance. Two parameters have to be chosen in RF regression: the number of candidate variables (mtry) to choose from at any split in the regression trees, and the number of trees (ntree). The number of variables to choose from was optimized by cross validation. The number of trees was fixed at 500 trees.

2.5 Criteria for comparison of the methods

2.5.1 Double cross validation

All methods above require input values for one or more hyperparameters (*e.g*. the number of components in PCR and PLS, the penalty parameter lambda in ridge regression and LASSO, etc.) and the values for these hyperparameters were optimized using cross validation. Using a single cross validation to estimate both the hyperparameters and the prediction error will result in an overly optimistic estimate of the error rate value (Smit et al., 2007). Hence, a double cross validation scheme was used (Stone 1974; Hendriks et al., 2007; Varma and Simon 2006). We used tenfold double cross validation (Hastie et al., 2001) for choosing optimum values for the hyperparameters and to estimate prediction error. First, tenfold cross validation is performed and one tenth portion of the data is left out for estimation of the prediction error, this portion is called the outer test set. The remaining nine tenth portions is the outer training set. Another tenfold cross validation uses nine tenth portions of the outer training data set which then are called the inner training sets and one tenth portions which are called the inner test sets. The hyperparameters are chosen which give the lowest MSEP values on the inner test data. We run this procedure 100 times, each with different tenfold divisions and in each division prediction was done and averaged over the results from 100 runs to obtain results in Table 1. The same divisions were used for all regression methods.

2.5.2 Mean squared error of prediction (MSEP)

The mean squared error of prediction (MSEP) is frequently used to assess the performance of regressions (Stallard et al.,1996; Mevik and Cederkvis 2004). MSEP of a regression can be estimated by predicting the test data set and comparing the predicted response with the observed response of the test set samples. Often, a (large enough) independent test set is not available. In such situations, the MSEP has to be estimated from the test data in cross-validation. An estimate of the MSEP is obtained by averaging the squared prediction errors of the outer test samples. Mathematically, we can write

$$\text{MSEP} = (1/n)\sum_{i=1}^{n} (y_i - y_{\text{predicted}})^2$$

Where y and $y_{\text{predicted}}$ are the observed and predicted response values for the i th test sample, respectively. We calculated and compared the MSEP on outer test sets for all the regression methods to evaluate the different methods. We consider the lowest MSEP to correspond to a better predictive model.

2.5.3 Variable selection or ranking

Variable selection is defined as selecting subsets of variables that together have predictive power. LASSO, SPLS and EN are variable selection methods as they select a subset of the predictor variables. For the variable selection methods we investigated the numbers of variables and the identity of the variables which were selected by those methods. For the methods that do not include variable selection we can still rank the variables according to their estimated regression coefficients or variable importance measures. In case of RR, PLS, PCR, RF all the variables remain in the regression model. In case of SVM, we do not perform variable ranking or variable selection as we cannot estimate regression coefficients. We compare the ranking between these different methods and we compare the ranks in the ranking methods with the selection of variables in the variable selection methods.

2.5.4 Goodness of fit (R^2)

Goodness of fit (R^2) of statistical models is used to describe how well the predictions fit a set of observations. It is a measure for the proportion of variability in a data set that is accounted for by the statistical model. In our analysis, we use R^2 values to compare the methods. R^2 is calculated as the square of the Pearson correlation between observed and fitted values for training and test data set and is converted to a percentage. The usual R^2 from a linear regression is just a measure of goodness-of-fit of the data at hand (training data), and not for future predictions (test data). We calculated R^2 values both for training and for the cross-validation test data.

2.6 OmicsFusion web application

OmicsFusion is a web-based application written in Java EE 6 and Struts 2 and runs on a glassfish v3 application server. SQLite v3 (http://www.sqlite.org) is used as the backend database management system. Standardized excel sheets are used to upload data to OmicsFusion. The end-user can select one or several of the described methods for data analysis. A Oracle Grid Engine 6.2u5-1 cluster (http://www.oracle.com) is used to execute the R based (http://cran.r-project.org/) script in parallel. The end-user is notified by email upon completion of the analysis. Results are summarized within the web-based interface. Results which are found in this paper by analyzing metabolomics data can be found in OmicsFusion with the identifier d8933.

3 Results

3.1 Univariate regression

Univariate regression analysis without FDR correction resulted in 29 significant variables ($p < 0.05$). MSEP and R^2 (training) values were calculated. Univariate regression followed by an FDR adjustment according to Benjamini and Hochberg (1995) resulted in still 23 significant variables at an FDR threshold of 0.05. Variable 294_0182 had the highest R^2 (training) of 22.6 % and lowest MSEP value of 1.92.

3.2 Comparison of the multivariate regression methods

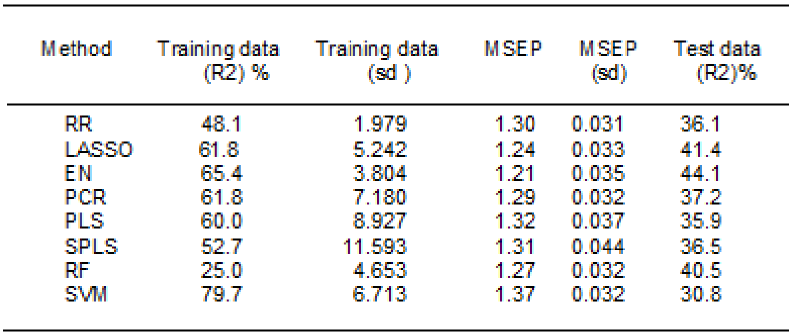
We used a double cross-validation scheme for comparison among the methods in 100 runs with different divisions of the data. In the table 1, we listed the differences in MSEP values between different methods were small. EN had the lowest MSEP value, lower than RR, PLS, SPLS, LASSO, SVM and RF (Table 1). The standard deviation of the MSEP values was about 0.03 for all the methods except SPLS (Table 1). Although SVM showed the highest R^2 value for the training data sets, it did not perform well for prediction on the test data.

3.4 Comparison of standardized regression coefficient based on ranking

We ranked the variables based on the standardized regression coefficients in all regression methods and then compared them to the ranks in univariate regression (Fig.1). In Fig. 1, variables are in x-axis whereas standardized regression coefficients are plotted in y-axis. Some of the variables get a zero value for the regression coefficient for those methods that also do variable selection (LASSO, SPLS, EN). Variables that were selected in these methods mostly correspond to variables that also had the largest regression coefficients in the univariate regressions, for example: mass_scan, 294_0182 gets the largest negative regression coefficient for both the variable selection (LASSO, EN, SPLS) and the ranking methods (PCR, PLS, RR and RF) and also in univariate regression. We also compared the ranking of the LASSO selected variables (24 mass_scans) in the ranking methods (PCR, PLS, RR, RF) and the variable selection methods (EN, SPLS) (Table 2).

Pearson correlation coefficients of the twenty-four variables selected by LASSO and flesh colour were visualized in a heat map (Fig. 2). From the Fig. 2, we can see that, there were high correlations among some of the selected variables, for example between mass_scans 396_1508 and 193_1508 with a correlation coefficient of 0.86; between 373_1301 and 557_1301 with a correlation coefficient of 0.85; between 396_1508 and 373_1301 with a correlation coefficient of 0.65; between 396_1508 and 557_1301 with a correlation coefficient of 0.60.

Table 1 Comparison of the eight multivariate regression methods based on R^2 for training data, standard deviation of R^2 (sd) for training data, MSEP, sd of MSEP and R^2 for test data set. R^2 for training data, MSEP and R^2 for test data are the mean values of the double cross validation scheme for 10 different divisions with 100 runs and then averaged.



Method	Training data (R2) %	Training data (sd)	MSEP	MSEP (sd)	Test data (R2)%
RR	48.1	1.979	1.30	0.031	36.1
LASSO	61.8	5.242	1.24	0.033	41.4
EN	65.4	3.804	1.21	0.035	44.1
PCR	61.8	7.180	1.29	0.032	37.2
PLS	60.0	8.927	1.32	0.037	35.9
SPLS	52.7	11.593	1.31	0.044	36.5
RF	25.0	4.653	1.27	0.032	40.5
SVM	79.7	6.713	1.37	0.032	30.8

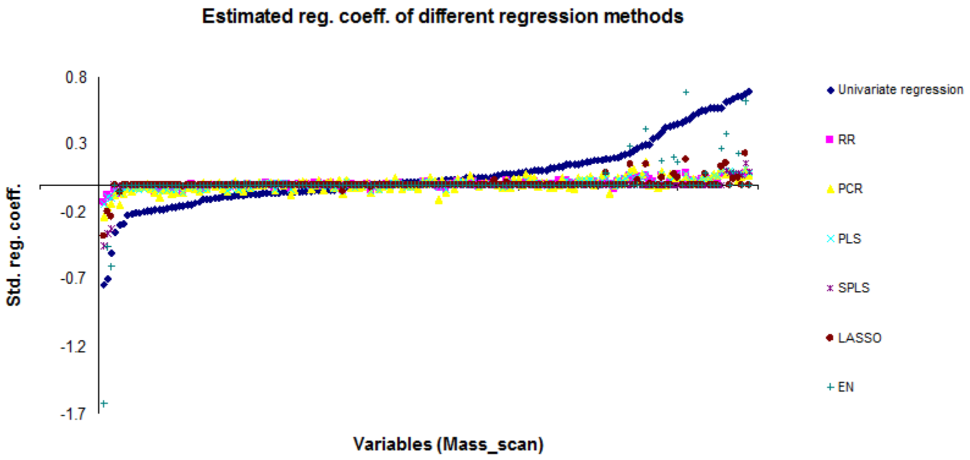


Figure 1: Standardized regression coefficients in the different regression methods. The order of the mass_scans on the x-axis is based on the regression coefficients from the univariate regressions. Variable 294_0182 has the highest negative regression coefficient in all the methods shown.

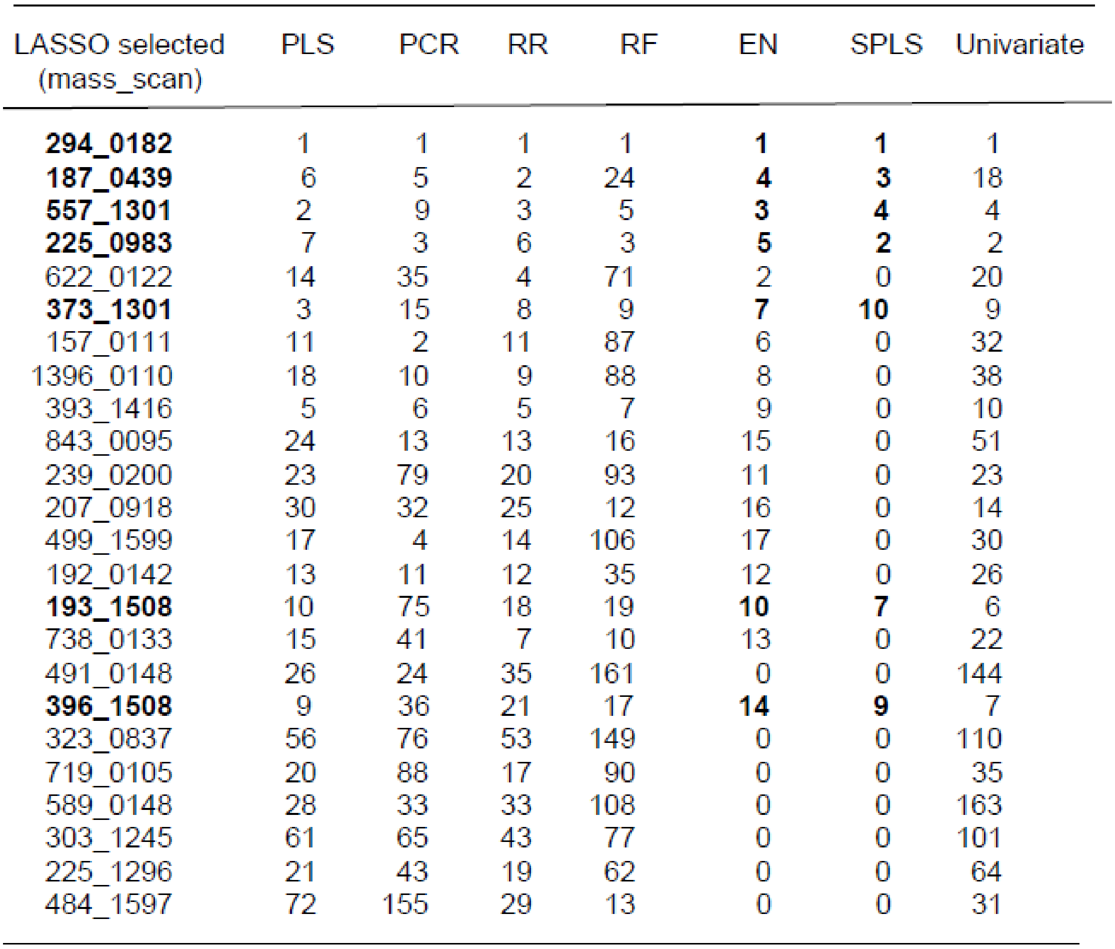
3.5 Comparison of standardized regression coefficients based on variable selection

We compared EN, LASSO and SPLS in terms of the numbers of selected variables. EN, LASSO and SPLS select 17, 24 and 10 variables, respectively. Seven variables are in common between all the three methods and shown in Fig. 5. For the pairwise comparison between EN and LASSO, 17 variables are in common. Between LASSO and SPLS, seven variables are in common. Between EN and SPLS seven variables are in common and all seven are also selected with LASSO.

3.6 OmicsFusion

The interface of OmicsFusion contains two major parts, namely: data submission and results visualization. Data submission allows end-users to start a new analysis in four distinctive steps: provide user details, upload of excel sheets, select analysis methods and start analysis after final confirmation. An unique token will be sent via email to the user. The OmicsFusion analysis for the data set described in this manuscript takes 38 minutes using 20 cores in parallel. Users are notified upon completion of the analysis via e-mail. The results of each analysis can be obtained after entering the unique token. The results, of all selected methods, will be summarized in a table and a snapshot is shown in Fig. 7. An overall mean rank is calculated for each of the predictor variables and the resulting table is ordered accordingly. The rank for the individual methods can be obtained by hovering over the coefficients. To quickly scan the results, the background of each coefficient is colour coded blue (top ranks) to white (lowest ranks). Each predictor variable is hyperlinked and can be used to show the response variable vs. predictor variable for easy interpretation of the results and shown in Fig. 8. This tool can be found: URL: http://www.plantbreeding.wur.nl/omicsFusion/

Table 2 Ranking of the twenty-four variables (mass_scans) selected by LASSO is shown in the first column. The other columns show the ranks of these twenty-four variables in PLS, PCR, RR, RF, EN, SPLS and univariate regression analysis. In RF ranking was done based on increase in MSE of the OOB samples after permutation of the variable. Out of these twenty-four variables, some are not selected by EN (17 variables selected) and SPLS (10 variables selected) and these are marked as 0. Variables in bold font are also selected in EN and SPLS.



LASSO selected (mass_scan)	PLS	PCR	RR	RF	EN	SPLS	Univariate
294_0182	1	1	1	1	**1**	**1**	1
187_0439	6	5	2	24	**4**	**3**	18
557_1301	2	9	3	5	**3**	**4**	4
225_0983	7	3	6	3	**5**	**2**	2
622_0122	14	35	4	71	2	0	20
373_1301	3	15	8	9	**7**	**10**	9
157_0111	11	2	11	87	6	0	32
1396_0110	18	10	9	88	8	0	38
393_1416	5	6	5	7	9	0	10
843_0095	24	13	13	16	15	0	51
239_0200	23	79	20	93	11	0	23
207_0918	30	32	25	12	16	0	14
499_1599	17	4	14	106	17	0	30
192_0142	13	11	12	35	12	0	26
193_1508	10	75	18	19	**10**	**7**	6
738_0133	15	41	7	10	13	0	22
491_0148	26	24	35	161	0	0	144
396_1508	9	36	21	17	**14**	**9**	7
323_0837	56	76	53	149	0	0	110
719_0105	20	88	17	90	0	0	35
589_0148	28	33	33	108	0	0	163
303_1245	61	65	43	77	0	0	101
225_1296	21	43	19	62	0	0	64
484_1597	72	155	29	13	0	0	31

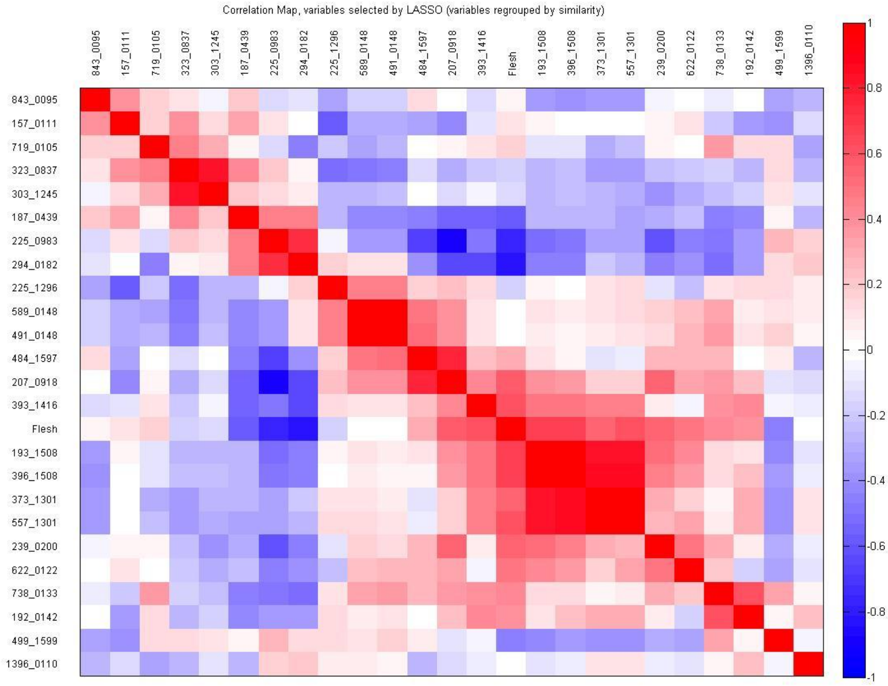


Figure 2: Correlation heat map of 24 variables selected with LASSO. Positive correlations among variables are shown in red, negative correlations are shown in blue. Variables 396_1508, 193_1508, 373_1301 and 557_1301 are grouped together based on high positive correlations.

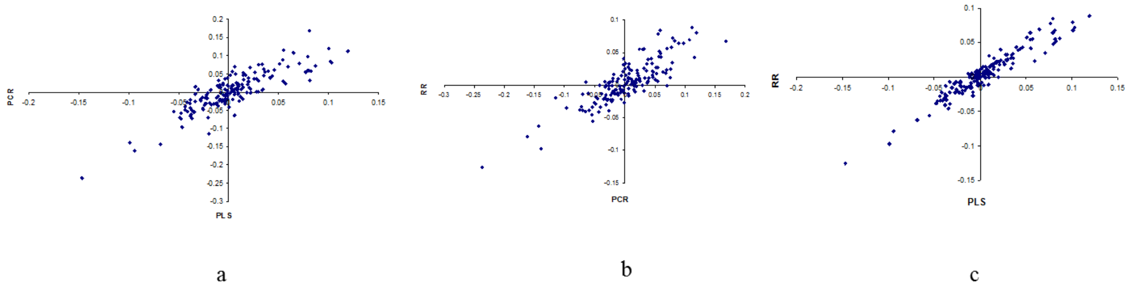


Figure 3: Comparison of standardized regression coefficients among PCR, PLS and RR a) PCR vs. PLS b) RR vs. PCR c) RR vs. PLS

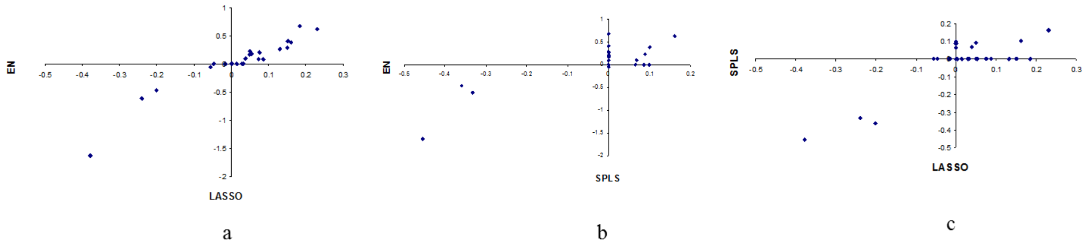


Figure 4: Comparison of standardized regression coefficients plot among LASSO, EN and SPLS. a) EN vs. LASSO b) EN vs. SPLS c) SPLS vs. LASSO

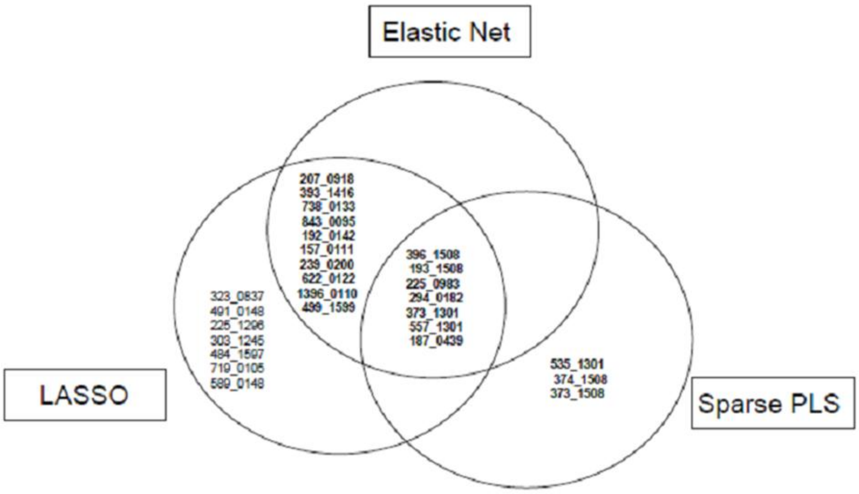


Figure 5: Venn diagram showing the overlap in selected variables among three penalized regression methods that include variable selection (EN, SPLS and LASSO). Seven variables are in common between all three methods.

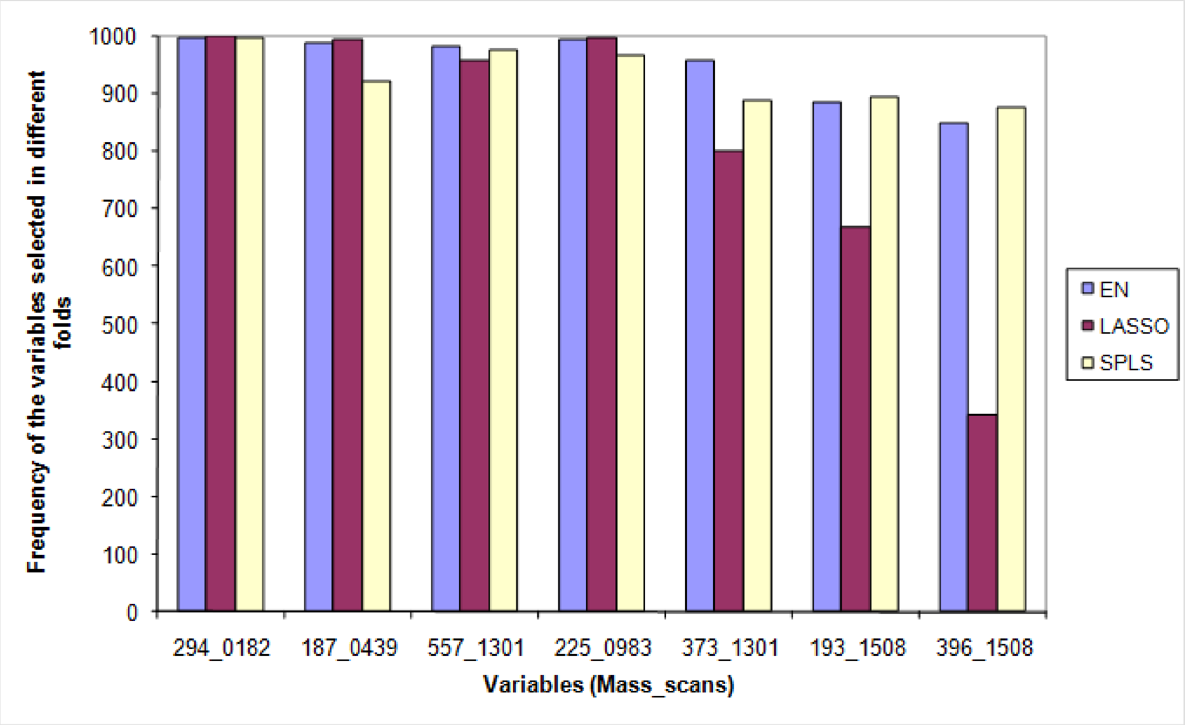
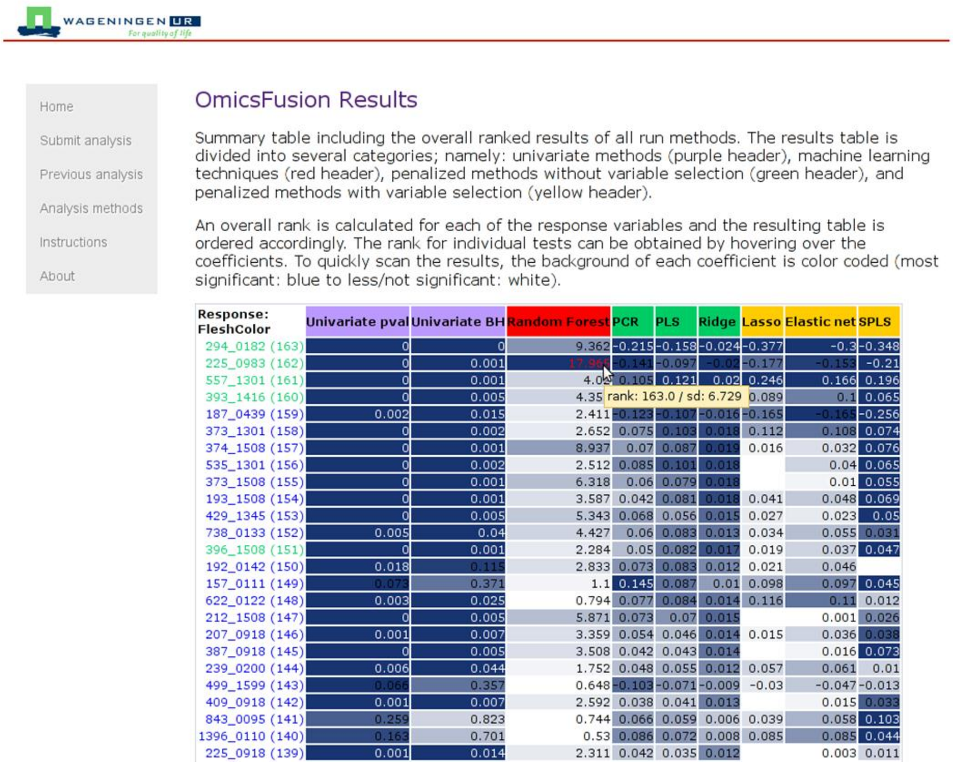


Figure 6: Frequency of the seven variables 294_0182, 187_0439, 557_1301, 225_0983, 373_1301, 193_1508 and 396_1508 over different folds in EN, LASSO and SPLS are shown. Variable 294_0182 shows the highest number of times selected in all the methods where as 396_1508 selected lowest numbers of times.



OmicsFusion Results

Summary table including the overall ranked results of all run methods. The results table is divided into several categories; namely: univariate methods (purple header), machine learning techniques (red header), penalized methods without variable selection (green header), and penalized methods with variable selection (yellow header).

An overall rank is calculated for each of the response variables and the resulting table is ordered accordingly. The rank for individual tests can be obtained by hovering over the coefficients. To quickly scan the results, the background of each coefficient is color coded (most significant: blue to less/not significant: white).

Response: FleshColor	Univariate pval	Univariate BH	Random Forest	PCR	PLS	Ridge	Lasso	Elastic net	SPLS
294_0182 (163)	0	0	9.362	-0.215	-0.158	-0.024	-0.377	-0.3	-0.348
225_0983 (162)	0	0.001	17.965	-0.141	-0.097	-0.02	-0.177	-0.153	-0.21
557_1301 (161)	0	0.001	4.02	0.105	0.121	0.02	0.246	0.166	0.196
393_1416 (160)	0	0.005	4.35	rank: 163.0 / sd: 6.729			0.089	0.1	0.065
187_0439 (159)	0.002	0.015	2.411	-0.123	-0.107	-0.016	-0.165	-0.165	-0.256
373_1301 (158)	0	0.002	2.652	0.075	0.103	0.018	0.112	0.108	0.074
374_1508 (157)	0	0.001	8.937	0.07	0.087	0.019	0.016	0.032	0.076
535_1301 (156)	0	0.002	2.512	0.085	0.101	0.018		0.04	0.065
373_1508 (155)	0	0.001	6.318	0.06	0.079	0.018		0.01	0.055
193_1508 (154)	0	0.001	3.587	0.042	0.081	0.018	0.041	0.048	0.069
429_1345 (153)	0	0.005	5.343	0.068	0.056	0.015	0.027	0.023	0.05
738_0133 (152)	0.005	0.04	4.427	0.06	0.083	0.013	0.034	0.055	0.031
396_1508 (151)	0	0.001	2.284	0.05	0.082	0.017	0.019	0.037	0.047
192_0142 (150)	0.018	0.115	2.833	0.073	0.083	0.012	0.021	0.046	
157_0111 (149)	0.073	0.371	1.1	0.145	0.087	0.01	0.098	0.097	0.045
622_0122 (148)	0.003	0.025	0.794	0.077	0.084	0.014	0.116	0.11	0.012
212_1508 (147)	0	0.005	5.871	0.073	0.07	0.015		0.001	0.026
207_0918 (146)	0.001	0.007	3.359	0.054	0.046	0.014	0.015	0.036	0.038
387_0918 (145)	0	0.005	3.508	0.042	0.043	0.014		0.016	0.073
239_0200 (144)	0.006	0.044	1.752	0.048	0.055	0.012	0.057	0.061	0.01
499_1599 (143)	0.066	0.357	0.648	-0.103	-0.071	-0.009	-0.03	-0.047	-0.013
409_0918 (142)	0.001	0.007	2.592	0.038	0.041	0.013		0.015	0.033
843_0095 (141)	0.259	0.823	0.744	0.066	0.059	0.006	0.039	0.058	0.103
1396_0110 (140)	0.163	0.701	0.53	0.086	0.072	0.008	0.085	0.085	0.044
225_0918 (139)	0.001	0.014	2.311	0.042	0.035	0.012		0.003	0.011

Figure 7: Summary of the OmicsFusion analysis. The overall rank is calculated for each of the predictor variables and the resulting table is ordered accordingly. The rank for the individual methods can be obtained by hovering over the coefficients. To quickly scan the results, the background of each coefficient is colour coded blue (top ranks) to white (lowest ranks).

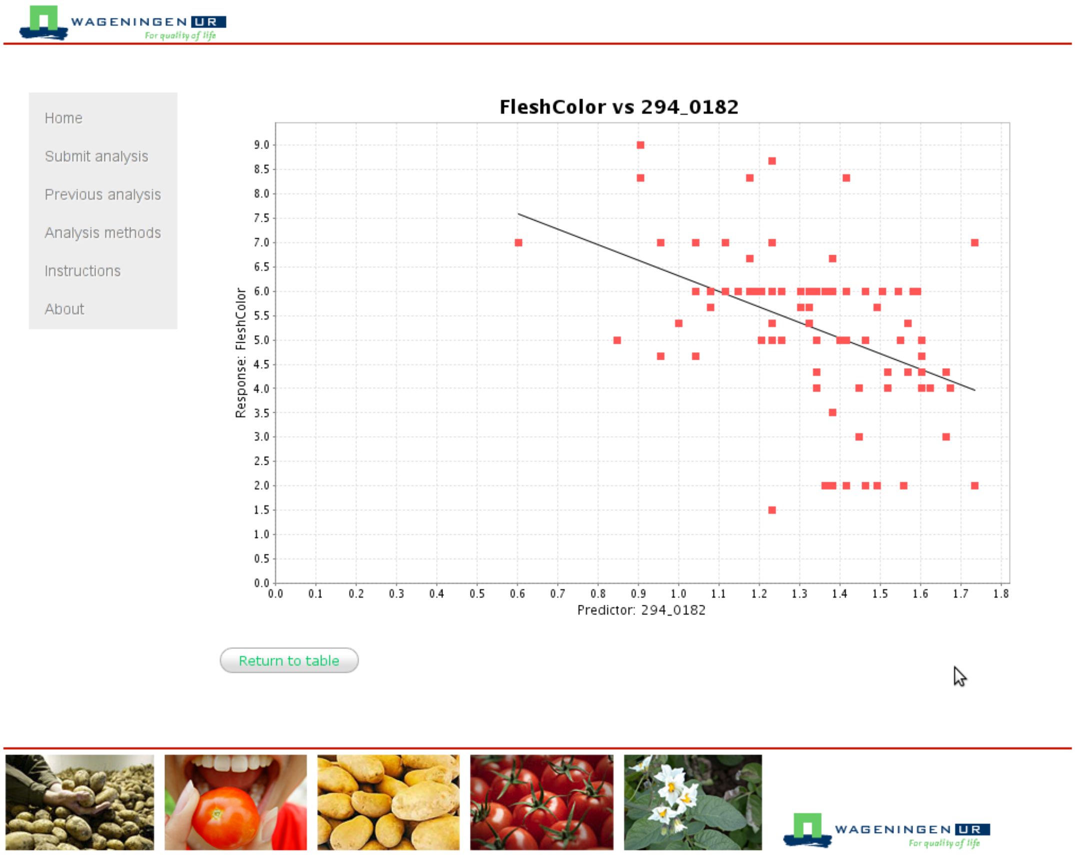


Figure 8: XY plot of the response variable tuber flesh colour vs. the predictor variable 294_0182. OmicsFusion can generate these XY plots for easy interpretation for each response vs predictor variable on-the-fly.

4 Discussion

We compared nine regression methods based on MSEP, R^2 on the training and on the test set, variable selection and variable ranking. The range of R^2 values for the training set across different methods is from 48.1% to 79.7%. R^2 is always lower on the test than on the training data, except for RF. RF includes an internal cross validation using the out-of-bag (OOB) samples. In addition, we used a double cross validation scheme for RF and the other multivariate methods. As a consequence the RF model is actually based on fewer samples than the other methods and therefore the R^2 value might be lower.

In the case of other methods there is a difference between the R^2 for the training and the R^2 for the test set. Taking the R^2 training as a criterion for evaluation of the methods to be used in multivariate regression of ~omics data is of limited use because it only refers to the fit of the data at hand and would be too optimistic for prediction purposes. Therefore we need instead to have a look at the MSEP. This is why we performed a double cross validation. EN has the lowest MSEP value which means that EN finds a better predictive model than other methods like RR, LASSO, SVM, PCR, PLS, RF and SPLS. RR, PCR and PLS have similar MSEP values. The lowest univariate MSEP was 1.92 for the variable 294_0182 but for the multivariate methods the MSEP values are lower, which suggests that these methods are predicting better than the best univariate predictor.

Comparing the variable selection methods, we saw that LASSO selected 24 variables and EN selects 17, a subset of those selected by LASSO. The average number of selected variables over 100 runs for EN, LASSO and SPLS were 31.3, 16.1 and 26.7 with standard deviations 11.9, 5.4 and 23.2, respectively. SPLS showed a high variability in the number of selected variables whereas LASSO had the lowest variability across 100 runs. Regarding consistency of the selected seven variables (Fig. 6) across 100 runs, we observed that variable 294_0182 was selected the highest number of times, almost always, in the different folds, whereas 396_1508 was selected the smallest number of times in EN, LASSO and SPLS.

Variables 294_0182, 187_0439 and 225_0983 are consistent in terms of the sign and size of standardized regression coefficients, ranking and size of standardized regression coefficients across methods (Fig. 4).

According to Tibshirani (1996) LASSO tries to select only one variable from a set of correlated variables but in our analysis we find that a group of correlated variables was selected, for example: mass_scans 396_1508 and 193_1508 have a correlation coefficient of 0.86; 373_1301 and 557_1301 have a correlation coefficient of 0.85 (Fig. 2). In addition, we did a simulation study which showed that only in sets of absolutely correlated variables (correlation coefficient of 1) LASSO picks one of the set of correlated variables and in that case the regression coefficient of the selected variable was close to double the simulated regression coefficient. In other situations it is possible that all or more of the correlated variables are selected together.

Regarding the grouping of correlated variables: five variables 373_1508, 396_1508, 374_1508, 193_1508 and 212_1508 are correlated with different correlation coefficients. Among these, 373_1508 and 374_1508 had a correlation coefficient of

0.97 (highest) and 193_1508 and 212_1508 a correlation coefficient of 0.74 (lowest). EN selects 396_1508 and 193_1508 with a correlation coefficient of 0.86 whereas SPLS selects 373_1508, 396_1508, 374_1508 and 193_1508.

The variables 535_1301, 373_1301 and 557_1301 are correlated with the highest correlation coefficient (0.94) between 535_1301 and 373_1301 and the lowest correlation coefficient (0.84) between 373_1301 and 557_1301. EN selects only two (373_1301 and 557_1301) whereas SPLS selects three of them.

SPLS performs better for selecting groups of correlated variables when compared to EN in selecting a larger number of correlated variables (simulation results, not shown).

Regarding the ranking methods (RR, PCR, PLS and RF): mass_scan 294_182 showed the highest absolute standardized regression coefficient in the different methods used and also the highest variable importance in RF. LASSO selected 24 variables which were ranked in decreasing order of the absolute standardized regression coefficients (Table 2). Within these 24 variables, the top 18 from PLS, top 12 from PCR, top 18 from RR, top 12 from RF, top 17 from EN, top 7 from SPLS and top 12 from univariate regressions were included. Variables like 557_1301 and 225_0983 obtained high ranks in all methods. Standardized regression coefficients of PLS, PCR and RR were more or less similar (Fig. 3). The standardized regression coefficients of RR and PLS are more similar than those regression coefficients of PCR. The correlation coefficient of standardized regression coefficients between PCR and PLS is 0.85, between RR and PCR is 0.85, between RR and PLS 0.95. These results confirm the observation of Hastie et al., 2001, in saying that “PLS, PCR and RR tend to behave similarly”.

Variable selection methods rather than non-selection methods here performed better in terms of the MSEP. This could be due to the fact that the variables which are not associated with the trait (noise variables) get regression coefficients with the value zero, so that they effectively drop out of the regression model. OmicsFusion offers an intuitive web-based interface which allows non-statisticians to easily analyze their own data using the statistical approaches described in this manuscript. In addition, OmicsFusion allows end-users to analyze data with more-than-one approach and summarizes the results of each method in an easy-to-interpret table. So, as an end user it serves as a web based omics analysis tool which produces results in terms of ranking and selection among the ~omics variables for prediction of a phenotypic trait of interest, using a variety of different methods, and it provides an easy summary of the most important results. The results table can be exported to Excel for further analysis and visualization.

In this paper, we have applied regression methods relating a phenotypic trait of interest as the response with a metabolomics data set, but the same methodology can be used in prediction of quantitative variables from other ~omics data sets as transcriptomics or proteomics data where also the numbers of samples (n) is usually much smaller than the number of variables (p). In addition, these prediction methods can also be applied in the context of genomic selection (Meuwissen et al., 2001) where prediction of phenotype is done from large data sets of molecular markers (Heslot et al., 2012).

Acknowledgements

The authors are grateful to Prof.Cajo ter Braak, Biometris, Wageningen UR, for his helpful discussion and valuable input.

Chapter 3

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

Animesh Acharjee[1,2], Bjorn Kloosterman[2], Ric C.H de Vos[3,4], Jeroen S. Werij[2,3], Christian W.B. Bachem[2], Richard G.F. Visser[2,3], Chris Maliepaard[2]

[1]Graduate School Experimental Plant Sciences

[2]Wageningen UR Plant Breeding, Wageningen University and Research Center, PO Box 386, 6700 AJ Wageningen, The Netherlands

[3]Centre for BioSystems Genomics, P.O. Box 98, 6700 AA, Wageningen, The Netherlands,

[4]Plant Research International, P.O. Box 16, 6700 AA Wageningen, The Netherlands

Published in Analytica Chimica Acta (2011) 705:56-63

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 colour and enzymatic discoloration. Network reconstruction has led to the integration of known and uncharacterized metabolites with genes associated to 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.

Key words: data integration, random forest, network reconstruction, tuber flesh colour, potato

1 Introduction

High-throughput ~omics technologies like microarrays (Brazma and Vilo 2000; Gaasterland and Bekiranov 2000) mass spectrometry (LC-MS, GC-MS) and protein chips Aebersold and Mann 2003; Zhu et al*.,* 2003) have gained much interest in the biological domain(Yuan et al., 2008). 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 (Joyce and Palsson 2006). 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" (Kitano 2002). 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 (Kim et al*.,* 2010; Fukushima et al*.,* 2009)

Several attempts have been made to integrate multiple ~omics data sets from different species such as metabolomics and proteomics (Wienkoop et al., 2008) and transcriptomics and metabolomics in *Arabidopsis thaliana*
(Bylesjo et al., 2007), transcriptomics and proteomics in soybean (Delmotte et al,. 2010) and transcriptomics, metabolomics and proteomics in grapevine berry (Zamboni et al., 2010).

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 to 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 the be the primary determinant of tuber flesh colour (Brown et al., 1993) and in recent years, much progress has been made in the identification of key regulatory genes in potato tuber carotenoid content (Campbell et al*.,* 2010; Wolters et al., 2010). A second tuber quality trait important for consumers and processing industry is enzymatic discoloration. Quantitative trait loci (QTLs) (Collard et al., 2005) for flesh colour and enzymatic discoloration and other quality traits have been reported (Wolters et al., 2010;Werij et al., 2007). The QTL analysis of both flesh colour and enzymatic discoloration within a well studied diploid potato population (here denoted as C x E) shows both unique and overlapping QTLs across the genome. Most interesting is the strong overlap between a QTL for flesh colour 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 CxE 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 colour. 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, also novel metabolites were found to be highly correlated to the phenotypic traits.

2 Material and methods

2.1 Plant material

Ninety-six individuals, including the parental clones, of a diploid backcross population (CxE) 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., 2003). 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 colour and determination of carotenoids are described in Kloosterman et al., 2010. Enzymatic discoloration (ED) after 5 minutes, 30 minutes, 3 hours and difference in discoloration between 3 hours and 30 minutes were described in Werij et al., 2007 . The Pearson correlation coefficient between enzymatic discoloration after 5 minutes and 30 minutes was very high (0.99), so out of these two we considered only enzymatic discoloration after 5 minutes.

2.2 Microarray hybridizations and data processing

RNA was extracted from the 96 samples using the hot phenol method described previously (Bachem et al., 1996). All samples were labeled with both Cy3 and Cy5-

dye using the low RNA input linear Amplification Kit, PLUS, Two colour (Agilent technologies) according to the manufacturer's protocol starting with 2 µg of purified total RNA. Hybridization and washing were performed according to the Agilent's two-colour 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-colour 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.

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., 2007. 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 700µl 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 (2.0x 150 mm, 3 mm particle size) using a 45 minutes 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 220-600nm 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 3,024 mass signals. Finally, the so-called multivariate mass spectra reconstruction strategy(Tikunov et al., 2005) 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., 2010. 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 x 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 (Kloosterman et al., 2010). For some of the carotenoids such as zeaxanthin (zea), violaxanthin (vio) and violaxanthin-like (vio-like) (Kloosterman et al., 2010). 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 by 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.

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/) (Moco et al., 2006), 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 (Breiman 2001; Diaz-Uriarte and Andres 2006) is a collection of unpruned decision trees (Hastie et al., 2001), used mostly for statistical classification but this method can be applied for regression as well (Segal 2004). 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 (Pang et al., 2006). 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 (Montgomery and Peck 1992). 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 (Breiman 2001). The greater the increase in the node purity values (Breiman 2001) the greater the importance of that particular variable (Breiman 2001). 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 (15062/3).

We used Random Forest (RF) for regression of the phenotypic traits flesh colour 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 colour 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 colour) 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 colour and enzymatic discoloration on gene expression values (Cy3 and Cy5 intensities) and metabolites separately were conducted by using the “Random Forest” package of R statistical software .

For transcriptomics (both for Cy3 and Cy5) and metabolomics analysis with flesh colour 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 (Fu et al., 2007). Separate linkage maps for “C” (Cmap) and “E” (Emap) parent were used (Bonierbale et al.,1988) for QTL analysis.

2.6 Network reconstruction

A network (Yuan et al., 2008) 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 threshold (α=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 (Batagelj and Pajek 2003). The schematic diagram of the methodology is shown in Fig. 1

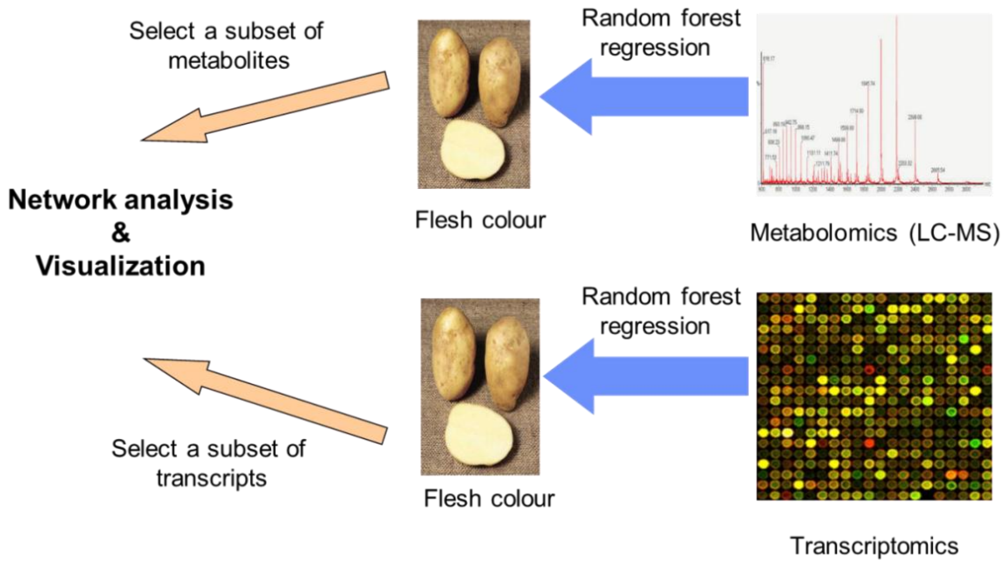


Figure 1: Integration of metabolomics and transcriptomics data with a trait (potato flesh colour or enzymatic discoloration) to identify important metabolites and genes associated with these traits, followed by a correlation network analysis and visualization.

3 Results

3.1 RF regression on the transcriptomics data set

Flesh colour

Random forest regression was applied to the transcriptomics data set using flesh colour as the response. The variance in flesh colour explained by the RF using Cy3 gene expression (R^2) 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 data 3). Based on our current knowledge of the potato tuber flesh colour or carotenoid content (Kloostermann et al., 2010) these genes were expected to be associated with flesh colour.

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 colour. In the top 50 of genes ranked by variable importance, beta-carotene hydroxylase (Bch) ranks first again, while zeaxanthin 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 R2 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.

Enzymatic discoloration

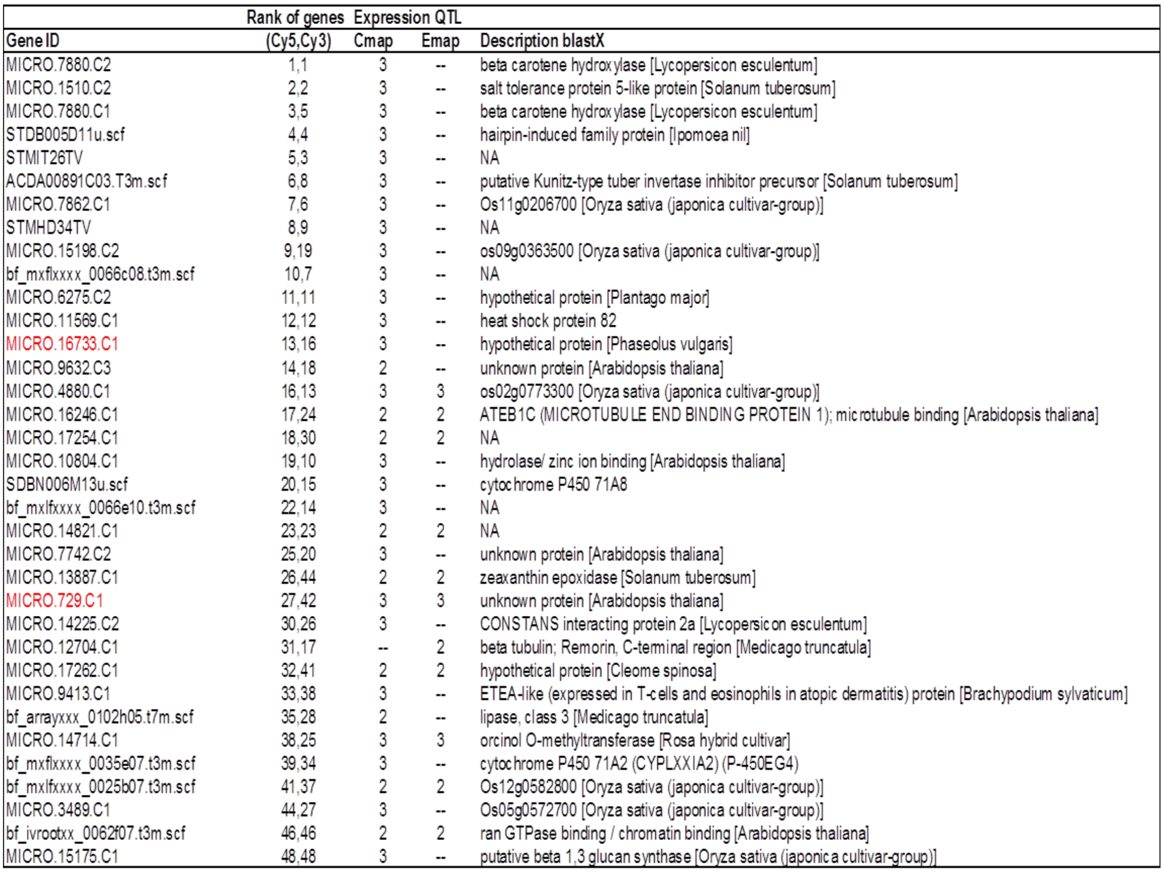
Enzymatic discoloration measured at different time points was also regressed on the transcriptomics data sets (Cy3 and Cy5, separately). Only enzymatic discoloration after 5 minutes had a positive value for the variance explained by the RF model (Table 3).

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 R2 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 colour and enzymatic discoloration after 5 minutes, 2 genes (Gene IDs: MICRO.16733.C1 and MICRO.729.C1) were in common (Table 1). The variance explained by a RF model using only the 12 genes in Cy3 and Cy5 was 48% and 46 % respectively.

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



	Rank of genes	Expression QTL		
Gene ID	(Cy5,Cy3)	Cmap	Emap	Description blastX
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)]

3.2 RF regression on the metabolomics data set

Flesh colour

We also applied RF to the metabolomics data set obtained from the same material used for expression analysis. The variance of flesh colour explained by the RF model for metabolites was 63%. Seven metabolites were significant (permutation threshold 0.001) (Fig. 2), the R2 value was also significant at this level. Using only these seven metabolites the variance explained by the RF model was 77 %.

Enzymatic discoloration

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

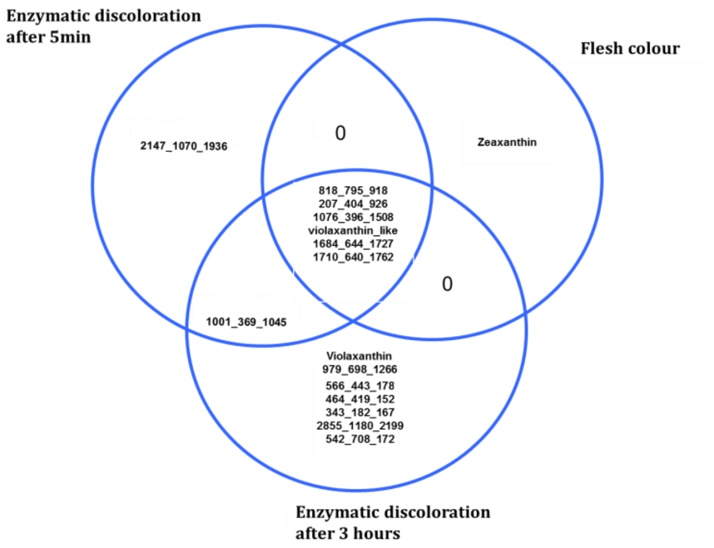
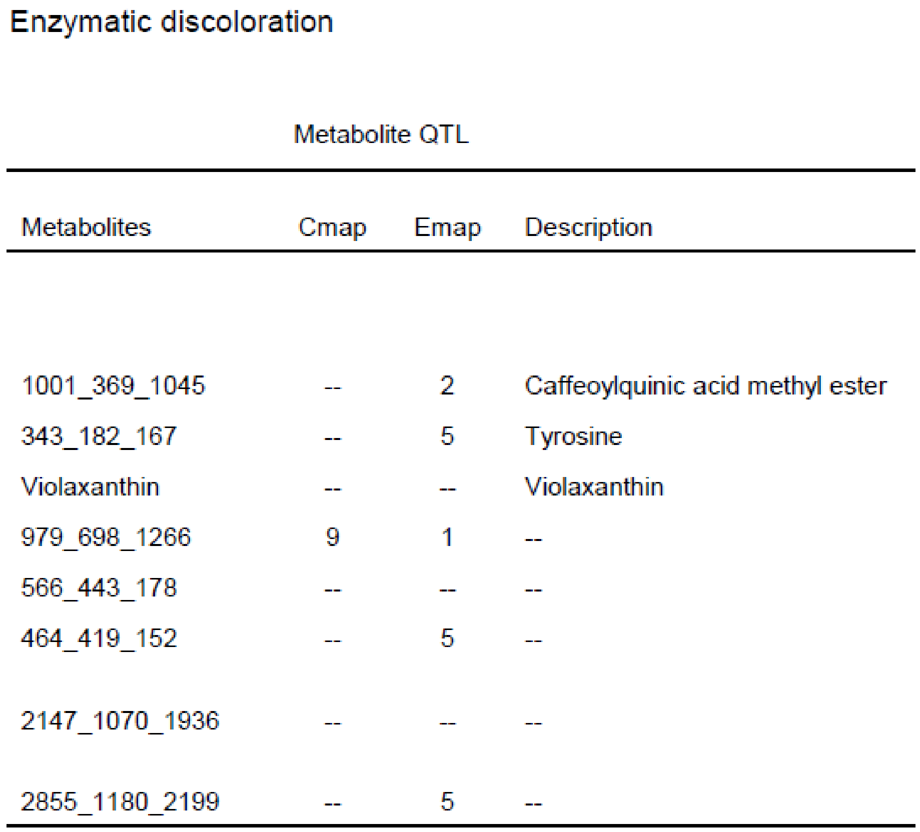


Figure 2: Venn diagram of metabolites associated with flesh colour, enzymatic discoloration after 5 min. Six metabolites are in common.

Table 2 Associated metabolites and mQTL analysis after RF regression for flesh colour and enzymatic discoloration

Flesh colour

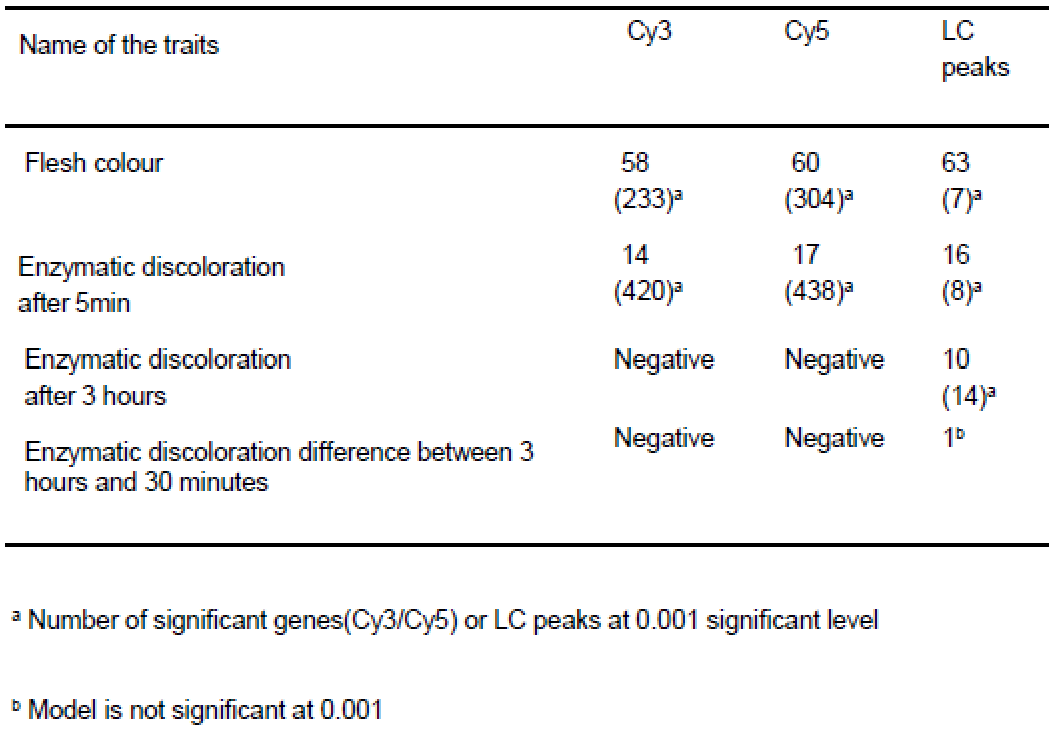
	Metabolite QTL		
Metabolites	Cmap	Emap	Description
818_795_918	3	--	2,3-Dihydroxy-4-megastigmen-9-one Glucoside
1076_396_1508	3	--	4,7-Megastigmadiene-3,9-diol Glucoside
207_404_926	3	2	--
1684_644_1727	3	3	--
1710_640_1762	3	--	--
Violaxanthin_like	3	--	Violaxanthin_like
Zeaxanthin	2	--	Zeaxanthin



Enzymatic discoloration

	Metabolite QTL		
Metabolites	Cmap	Emap	Description
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 colour	58 (233)[a]	60 (304)[a]	63 (7)[a]
Enzymatic discoloration after 5min	14 (420)[a]	17 (438)[a]	16 (8)[a]
Enzymatic discoloration after 3 hours	Negative	Negative	10 (14)[a]
Enzymatic discoloration difference between 3 hours and 30 minutes	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

3.3 Integration of transcriptomics and metabolomics data

For flesh colour, we used the 35 significant genes in common in the top 50s 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 (3h and difference 3h-30min), we did not get any significant RF regression model (R2 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 colour, 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 colour and are included (Fig. 2).

For flesh colour, we took 2 significant genes (Bch, Zep) from the transcriptomics data set which were already reported in the literature(Brown et al., 2006; Wolters et al.,

2010) 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 2 genes and significant metabolites.

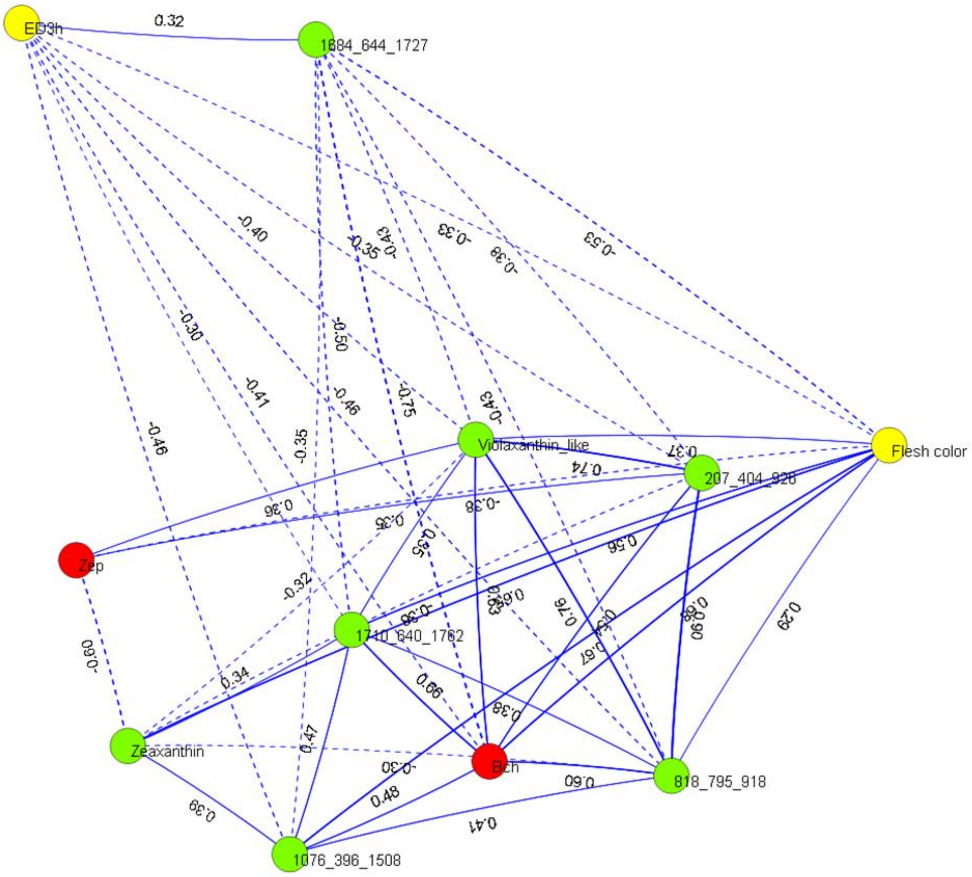


Figure 3: Pearson correlation network of genes (red), metabolites (green) and phenotypic traits (yellow). The dotted lines represent negative correlation coefficients, solid lines represent positive correlation coefficients. Only correlation coefficients significant at p<0.01 are considered. Bch = Beta-carotene hydroxylase, Zep= Zeaxanthin epoxidase, ED3h= Enzymatic discoloration after 3 hours.

4 Discussion

Flesh colour

We used RF regression for integrating transcriptomics and metabolomics data sets with potato tuber flesh colour. In the trancriptomics 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)(Brown et al., 2006) and zeaxanthin epoxidase (Zep)(Wolters et al., 2010). were ranked high, both of which have previously been associated with flesh colour in potato tubers (Brown et al., 2006; Wolters et al., 2010). 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(Kloostermann et al., 2010). 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 (Wolters et al., 2010). 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*.,* 2009) used for identification of epistatic interactions.

The majority of the additional genes with high ranking show eQTLs on chromosome 3 and 2 directly under the QTLs for flesh colour. If one considers the flesh colour 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 colour 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 colour we find a large representation of genes (24 genes) on chromosome 3 and 2 in the same regions as the flesh colour 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 (Stushnoff et al., 2010), has an eQTL coinciding with the QTL for flesh colour 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 (CCD’s) involved in carotenoid breakdown (Campbell et al., 2010). 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 colour. Based on their accurate masses and in-source fragmentation 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 glucosides 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 (Zamboni et al., 2010) 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.

Enzymatic discoloration: in the transcriptomics analysis only enzymatic discoloration after 5 minutes was significantly associated to 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 (Werij et al*.,* 2007) QTLs for enzymatic discoloration were mapped to chrom 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 to the remaining QTLs have not yet been identified. From the metabolomics analysis, out of 8 significant untargeted metabolites we putatively identified 2 metabolites as Caffeoylquinic acid methyl ester
and tyrosine, which were already reported by (Werij et al., 2007) and associated with enzymatic discoloration and confirm our findings. Other metabolites in table 2 (such as 2147_1070_1936, 979_698_1266 *etc*.) 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 colour and enzymatic discoloration (Fig. 2). QTL analysis of tuber flesh colour 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 (Rigal et al., 2000). 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 colour, 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 7 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 minutes, 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 whereas with 8 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 to 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 colour 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 colour, 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 colour per se, and the involvement of more genes in enzymatic discoloration than in flesh colour.

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 (Breiman 2001) (using the OOB samples) and has only a few tuning parameters which if chosen reasonably, do not change results strongly (Gislason et al., 2006). 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 (Steinfath et al., 2010). 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.

Acknowledgements

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

Supplementary Table 1, 2 and 3 can be found

http://www.sciencedirect.com/science/article/pii/S0003267011004508

Chapter 4

Untargeted metabolic quantitative trait loci (mQTL) analyses reveal a relationship between primary metabolism and potato tuber quality

Natalia Carreno-Quintero[1,2,*], Animesh Acharjee[3,*], Chris Maliepaard[3], Christian W.B. Bachem[3], Roland Mumm[2], Harro Bouwmeester[1,2] , Richard G.F. Visser[3], Joost J.B. Keurentjes[1,2,4]

[1]Laboratory of Plant Physiology, Wageningen University, 6708 PB Wageningen, The Netherlands

[2]Plant Research International, Wageningen UR, 6708PB Wageningen, The Netherlands, Centre for BioSystems Genomics, 6700 AB Wageningen, The Netherlands

[3]Wageningen UR Plant Breeding, Wageningen University, 6708 PB Wageningen, The Netherlands

[4]Laboratory of Genetics, Wageningen University, 6708 PB Wageningen, The Netherlands

*Equal contributions

Published in Plant Physiology (2012)158:1306-1318

Abstract

Recent advances in ~omics technologies such as transcriptomics, metabolomics and proteomics along with genotypic profiling have permitted to dissect the genetics of complex traits represented by molecular phenotypes in non-model species. To identify the genetic factors underlying variation in primary metabolism in potato we have profiled primary metabolite content in a diploid potato mapping population, derived from crosses between *Solanum tuberosum* and wild relatives, using gas chromatography time of flight mass spectrometry (GC-TOF-MS). In total 139 polar metabolites were detected of which we identified metabolite quantitative trait loci (mQTLs) for ~72% of the detected compounds. In order to obtain an insight into the relationships between metabolic traits and classical phenotypic traits we also analysed statistical associations between them. The combined analysis of genetic information through QTL coincidence and the application of statistical learning methods provide information on putative indicators associated with the alterations in metabolic networks that affect complex phenotypic traits.

Key words: GC-MS, mQTL, metabolomics

Introduction

The variation observed in phenotypic trait values in plants is often of quantitative nature and it remains challenging to unravel the genetic basis of these traits. Quantitative trait locus (QTL) mapping is currently the most commonly used approach to dissect the genetic factors underlying complex traits. The goal of QTL mapping is to identify genomic regions associated with a specific complex phenotype by statistical analysis of associations between genetic markers and phenotypic variation (Doerge 2002). Recently, advances in high-throughput analysis and analytical detection methods have facilitated more integrated approaches to measure global phenotypic variation at the molecular level. Metabolite profiling is a rapidly evolving technology which has significantly increased the possibilities of performing high-throughput analysis of hundreds to thousands of compounds in a range of plants including complex crop species. Metabolite composition is of great importance in crop plants as a number of important traits such as biotic and abiotic stress responses, post-harvest processing and nutritional value are largely dependent on the metabolic content (Fernie and Schauer 2009).

In potato breeding metabolomic studies have progressively increased in importance as many potato tuber traits such as content and quality of starch, chipping quality, flesh colour, taste and glycoalkaloid content have been shown to be linked to a wide range of metabolites (Coffin et al., 1987; Dobson et al., 2008). As a result, tuber quality can be assessed by assaying a range of metabolites. Gas chromatography (time of flight) mass spectrometry (GC-TOF-MS) has been shown to be useful for the rapid and highly sensitive detection of a large fraction of plant metabolites covering the central pathways of primary carbon metabolism (Roessner et al., 2000; Lisec et al., 2006). In potato, untargeted metabolomic approaches by GC-MS have been successfully applied to assess changes in metabolites under different conditions (Roessner et al., 2000; Urbanczyk-Wochniak et al., 2005), to evaluate the metabolic response to various genetic modifications (Roessner et al., 2001; Szopa et al., 2001; Davies et al., 2005), and to explore the phytochemical diversity among potato cultivars and landraces (Dobson et al., 2008; Dobson et al., 2009). Additionally, metabolite profiling has been used to follow changes during key stages in the tuber life cycle (Davies 2007). Untargeted approaches have thus generated a substantial amount of data providing important information concerning compositional metabolite changes upon perturbation and phytochemical diversity in potato. However, so far these studies have focused on applications of metabolite profiling as an evaluation and comparative tool. Technological developments and improved data processing techniques now also allow the use of metabolite profiling to obtain further insight into the genetic factors controlling metabolic traits (Keurentjes 2009). Exploration of the genetic factors underlying metabolite variation in mapping populations is particularly advantageous. The genetic variation between related individuals in a segregating mapping population can be exploited to locate the genomic regions involved in the regulation of the observed metabolite variation (Keurentjes et al., 2006).

Here, we report on the metabolic profiling of a segregating diploid potato population (C x E). This population is highly heterozygous and has been analysed in a number of studies to investigate the genetic architecture of quantitative traits (Van Eck et al., 1994; Werij et al., 2007; Kloosterman et al., 2010; Wolters et al., 2010). For a number of traits candidate genes and their allelic variants have been identified from these studies, including tuber flesh colour and cooking type (Kloosterman et al., 2010), tuber shape (Van Eck et al., 1994), carotenoid biosynthesis (Wolters et al., 2010) and methionine content (Kloosterman et al., 2010).

In this study we have used the C x E population to explore the genetic basis of primary metabolites. To this end an untargeted GC-TOF-MS metabolic profiling was carried out on a core set of individuals of the C x E population. In order to investigate the variation in the detected metabolite levels in the individuals of the population we applied a genetical genomics approach (Jansen and Nap 2001). QTL analysis of metabolite levels resulted in the identification of genomic regions associated with the metabolic variation. In addition, we performed a parallel QTL analysis for starch and cold sweetening related traits and genetically inferred links were established between these phenotypic traits and primary metabolites. We further applied multivariate analysis to the combined data sets of starch related traits and metabolic profiles to determine the predictive power of metabolites on a given phenotypic trait. For this we used a Random Forest (RF) (Breiman 2001) approach to find significant associations between phenotypic and metabolic traits independent of genetic information. Putative predictors were tested and confirmed in an independent collection of potato cultivars.

Our results show the value of combining biochemical profiling with genetic information to identify associations between metabolites and phenotypes. This approach reveals previously unknown links between phenotypic traits and metabolism and thus facilitates the discovery of biomarkers for agronomical important traits.

Materials and Methods

Plant material

The C x E population

The diploid population (C x E) consisting of a total of 251 individuals was obtained from a cross between two heterozygous diploid potato clones, USW5337.3 (coded C: *Solanum phureja x Solanum tuberosum*) and 77.2102.37 (coded E: *S.vernei x S.tuberosum*). The development of the popualtion is described in detail in (Celis-Gamboa 2002).

A subset of 97 genotypes of this population was grown in two subsequent years (2002 and 2003) during the normal potato growing season (April-September) in Wageningen, The Netherlands. For each genotype, tubers were collected from three plants. Harvested tubers were either used for phenotypic analysis or mechanically peeled and immediately frozen in liquid nitrogen before being ground into fine powder and stored at -80^{o}C.

Phenotypic analyses for 26 starch and cold sweetening related traits were performed for both years of harvest (Supplemental Table S6). Metabolite profiling was carried out on the ground material of tubers of the 2003 harvest.

Potato cultivars

Potato cultivars used for independent confirmation and further statistical analysis were part of the potato collection available at Wageningen UR Plant Breeding. This collection consists of 221 tetraploid potato cultivars which were provided by Dutch breeding companies and gene banks. Phosphate measurements were carried out for 214 potato cultivars. In accordance with the distribution of the trait values (Supplemental Fig. S2), we selected 30 cultivars representing high, medium and low phosphate content.

Determination of starch phosphorylation

The degree of phosphorylation of starch was determined in a colorimetric assay. 20 mg starch was mixed with 250 µl of 70% $HClO_4$ and incubated at 250°C for 25 min. Then 50µl of 30% H_2O_2 (w/v) was added and incubated for another 5 min. After cooling, 2ml of H_2O was added to reach a final concentration of $HClO_4$ 8.75% (w/v).

The colour reagent consisted of 0.75% (w/v) $(NH_4)6Mo_7O_{24}.4H_2O$, 3% (w/v) $FeSO_4.7H_2O$ and 0.75% SDS (Sodium Dodecyl Sulphate) (w/v) dissolved in 0.375 M H_2SO_4. 100 µl of the sample extract, or a standard solution, was mixed with 200 µl of the colour reagent solution in a microtiter plate and incubated for 10 minutes at room temperature. The absorbance was measured at 750 nm in a microplate reader using 8.75% $HClO_4$ as a blank. A calibration curve of PO_4 dissolved in $HClO_4$ (0 – 500 µM) was used to determine the phosphate content.

Extraction and derivatization of potato tuber metabolites for GC-MS analysis

Relative metabolite content was determined as described in (Weckwerth et al., 2004) with modifications specific to the potato material. Briefly, polar metabolite fractions were extracted from ~100 mg fresh weight (FW) of tuber powder. 1.4 ml of a single phase solvent mixture of methanol:chloroform:water (2.5:1:1) was added to the ground tuber powder in a 2ml eppendorf tube. Alanine-*d3* was used as deuterated internal standard and ribitol was used as a representative internal standard and they were all mixed in one solution. In the water phase: 25 ul of a solution containing the aforementioned internal standard solution was added. After vortexing, the closed tubes were sonicated for 15 min. After 5 minutes of centrifugation the supernatant was transferred into a new eppendorf tube and 400 µl of water was added. The mixture was thoroughly mixed by vortexing and centrifuged for 10 min at 21000 rfc. The methanol/water supernatant (polar phase) was carefully transferred into a new eppendorf tube. Aliquots of the polar phase (100 µl) were dried by vacuum centrifugation for 12-16 hours.

The dried samples were derivatized on-line as described by Lisec et al. 2006 using a Combi PAL autosampler (CTC Analytics AG). First, 12.5µL O-methlhydroxylamine hydrochloride (20 mg/ml pyridine) was added to the samples and incubated for 30 min at 40ºC with agitation. Then the samples were derivatized with 17.5µL MSTFA (N-methyl-N-trimethylsilyltrifluoroacetamide) for 60 min. An alkane mixture (C9-C17,C20-C34) was added to determine retention indices of metabolites. The derivatized samples were analysed by a gas chromatography-time of flight-mass spectrometry (GC-TOF-MS) system consisting of an Optic 3 high performance injector (ATAS GL Int.) and an Agilent 6890 gas chromatograph (Agilent Technologies) coupled to a Pegasus III time-of-flight mass spectrometer (Leco Instruments).

2 µl of each sample was introduced to the injector at 70^{o}C using a split flow of 19 ml/min.The injector was rapidly heated with 6 ^{o}C/s to 240^{o}C. The chromomatographic separation was performed using a VF-5ms capillary column (Varian; 30m x 0.25 mm x 0.25 µm) including a 10m guardian column with helium as carrier gas at a column flow rate of 1 ml/min. The temperature was isothermal for 2 min at 70^{o}C, followed by a 10^{o}C per minute ramp to 310^{o}C, and was held at this temperature for 5 min. The transfer line temperature was set at 270^{o}C. The column effluent was ionised by electron impact at 70eV. Mass spectra were acquired at 20 scans/sec within a mass range of m/z 50 – 600, at a source temperature of 200°C. A solvent delay of 295-s was set. The detector voltage was set to 1400V.

GC-MS data processing methods

Data pre-processing

Raw data were processed by ChromaTOF software 2.0 (Leco instruments) and MassLynx software (Waters Inc.) and further analysis was performed using MetAlign software (Lommen 2009) to extract and align the mass signals (s/n≥ 2). Mass signals that were present in less than 2 samples were discarded. Signal redundancy per metabolite was removed by means of clustering and mass spectra were reconstructed (Tikunov et al., 2005). This resulted in 139 reconstructed polar metabolites (representative masses).

Compound identification

The mass spectra of the representative masses were subjected to tentative identification by matching to the NIST05 (National Institute of Standards and Technology, Gaithersburg, MD, USA; http://www.nist.gov/srd/mslist.htm) and Wiley spectral libraries and by comparison with retention indices calculated using a series of alkanes and fitted with a second order polynomial function (Strehmel et al., 2008). Library hits were manually curated, and a series of commercial standards were used to check annotation. Compound identification is limited to the availability of spectra in the libraries used. The identities of the detected compunds are listed in Supplemental Table S1.

Data normalization and Multivariate Analysis

Mass intenstity values of the representative masses were normalized using isotope labeled *d3*-alanine as an internal standard. Relative amounts of the compounds were obtained by normalizing the intensity of individual masses to the response of the internal standard. The ratio between the mass intensity value of the putative compound and the *d3*-alanine internal standard was then scaled by multiplying the resulting value by the average of the *d3*-alanine mass intensity across all samples.

Normalized values were log-transformed in GeneMaths XT version 2.12 software (www.applied-maths.com). These data were used for cluster analysis using Pearson's correlation coefficient and UPGMA for hierachical clustering method.

Metabolic and phenotypic QTL analysis

Metabolite QTL analyses were performed using the software package Metanetwork (Keurentjes et al., 2006; Fu et al., 2007). Metanetwork applies a two-part model and a *p*-value is determined for each part of the model. In this study, *p*-values and QTL thresholds were determined as described in (Keurentjes et al., 2006). Since Metanetwork is not designed for cross-pollinated species, two separate linkage maps were used in our analysis: one for the female parent C and one for the male parent E. The number of markers specific to the C parent map is 218 and for the E parent map 178 with an average spacing between markers of 6.1 and 3.9 cM respectively. The significance QTL threshold value was estimated by Metanetwork. Empirical thresholds for significant mQTLs were calculated separately for both parental maps, C-parent map: $-{}^{10}\log(p) = 3.43$, (p=0.00037) and E-parent map: $-{}^{10}\log(p) = 3.19$, (p=0.00065).

Phenotypic measurements containing missing data cannot be analyzed by Metanetwork, hence, QTL analyses for phenotypic data were performed using the software package MapQTL® Version 6.0 (Van Ooijen 2009). QTL LOD thresholds were calculated per trait using a permutation test (N = 10,000) provided in MapQTL®.

Broad sense heritability was estimated for starch phosphorylation measurements over the two years (2002, 2003) according to the formula $H^2 = V_G / (V_G + V_E + V_G{*}_E)$, where V_G is the variance among genotypes and V_E *is* the year variation. One phosphate content measurement per year was used in a mixed model to calculate variance components for genotypes, years and residual (=genotype * year).

Random Forest

Random Forest (RF) (Breiman 2001) was used for regression of the phenotypic trait starch phosphorylation on the GC-TOF-MS signals. RF constructs a predictive model for the response using all predictors but quantifies the importance of each, here the metabolites, in explaining the variation present in the starch phosphorylation. RF by itself does not provide significance levels of individual metabolites and does not perform a variable selection to choose a possible subset of associated metabolites. Therefore, we included a permutation test to indicate significance of the association of a metabolite with a trait. In each of 1,000 permutations of the trait values we estimated the variance explained by the RF model (R2) and the variable importance of each metabolite in terms of the decrease in node impurities (Breiman 2001). We ordered node purity values from the permuted data sets and took the 95 percentile from the distribution of impurity values as the significance threshold of the individual metabolites. The same procedure was done for R2 values of the model: The 95 percentile was taken as a significance threshold for the RF model. RF regression of starch phosphorylation on metabolite values was conducted using the “randomForest” package of R statistical software. R2 in Random Forest is not just a measure of goodness-of-fit of the data at hand but is determined on left-out samples (the ‘out-of-bag’ samples) so it should be interpreted as a measure for predictive quality (here considered as prediction R2) of the Random Forest on independent samples that have the same properties as the in-bag samples. (Breiman 2001).

Results

Metabolite profiling

In order to assess the content and variation of polar primary metabolites present in the C x E diploid potato population an untargeted GC-TOF-MS-based metabolite profiling was performed. The GC-TOF-MS method was applied to the polar aqueous methanol extracts of dormant tubers of 97 genotypes and the parental clones of the C x E population. After processing of the raw data 139 representative masses were obtained. The distribution of trait values for the detected compounds across all the genotypes was wide, with coefficients of variation higher than 50% for the majority of metabolites (Fig 1). This large variation can in part be explained by the segregation of genetic factors and therefore is amenable to genetic mapping approaches.

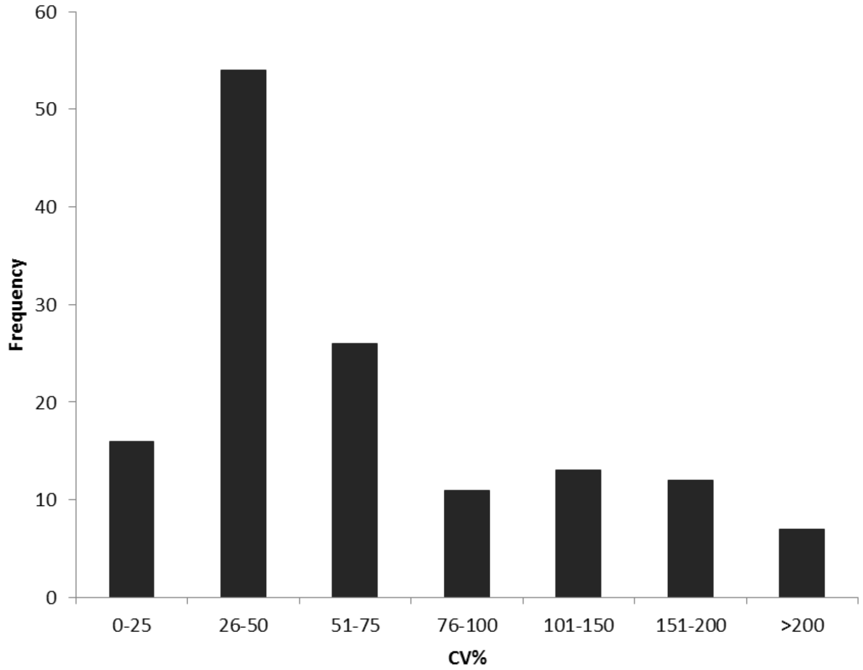


Figure 1: Histogram of the distribution of metabolic variation . Data shown are based on the CV% across the C x E population for the 139 representative masses. The distribution shows that the majority of metabolites have a CV higher than 50%, indicating a high level of variability.

The 139 representative masses were putatively identified on the basis of their fragment masses using NIST (http://www.nist.gov/srd/mslist.htm). Supplemental Table S1 lists the derived identification (i.e. putative metabolites). Further inspection of the spectra and retention indices of these fragments allowed a more accurate annotation of 58 of them (Table 1). Using all samples we performed a hierarchical cluster analysis using Pearson correlations on the processed data of the abundances of the 139 representative masses. Fig. 2 shows the degree of correlation among the detected compounds. The majority of the compounds were identified as amino acids, organic acids or carbohydrates. The annotation of a number of compounds could not be verified by further inspection of the spectra and retention indices. However, we included these unknown compounds in the cluster analysis to investigate the degree of association with identified metabolites. We found that compounds from the same class, such as amino acids or carbohydrates, generally clustered. The correlation coefficients within the identified amino acids ranged between 0.6 and 0.9 (Fig. 2) and two metabolites were considered to be highly correlated if the absolute correlation coefficient had a value ≥ 0.6. Such high correlations have been reported before in potato between amino acids (Roessner et al., 2001; Dobson et al., 2008). It has been suggested that this correlation may reflect the mechanism of general amino acid control in plants (Halford et al., 2004). Interestingly, within the amino acid cluster the branched amino acids isoleucine, leucine, and valine clustered separately from the aromatic amino acids tyrosine, phenylalanine and tryptophan, as was also reported in earlier metabolomics studies in different potato cultivars (Roessner et al., 2001; Noctor et al., 2002; Dobson et al., 2008). Amino acids that are biosynthetically linked such as serine, glycine, and threonine were also highly correlated (Fig. 2). This indicates that much of the variation in amino acid content is genetically controlled by a few master regulators. However, this was not the case for all related amino acids. The Pearson correlation coefficients among GABA (γ-aminobutyric acid), glutamic acid and proline were less than 0.2, although they are closely linked biosynthetically as members of the glutamate family. Other amino acids, such as glutamic acid and asparagine, show weak correlations (<0.4) with the major cluster of amino acids. This could suggest that the genetic regulation of these biosynthetic routes is independent from that of the cluster of amino acids. In addition, most of the detected sugars, such as mannose and fructose, also form a cluster (Fig. 2). In contrast, sucrose clusters with a group of organic acids rather than with other sugars. The clustering of organic acids is, however, less distinct, and this is likely to be due to the diverse biochemical origins of these compounds.

Identification of metabolic QTLs (mQTLs)

To determine if the variation observed in metabolite levels could indeed be explained by allelic differences in genetic factors, we performed metabolic QTL (mQTL) analyses on the metabolic profiles. The software package Metanetwork was used to map the metabolite variation. Metanetwork applications (Fu et al., 2007) were designed from data collected from recombinant inbreed lines (RILs), hence, in order to adjust the software applications to a cross-pollinated species like potato, we used two separate linkage maps: one for the female parent C and one for the male parent E. Overall, detected variation in abundance of approximately 72% of the metabolites could be mapped in at least one of the two linkage maps. In total, we found 187 mQTLs for 121 metabolites, of which 58 could be putatively annotated (Supplemental Table S1). A complete list and description of the detected mQTLs is given in Supplemental Tables S2 and S3.

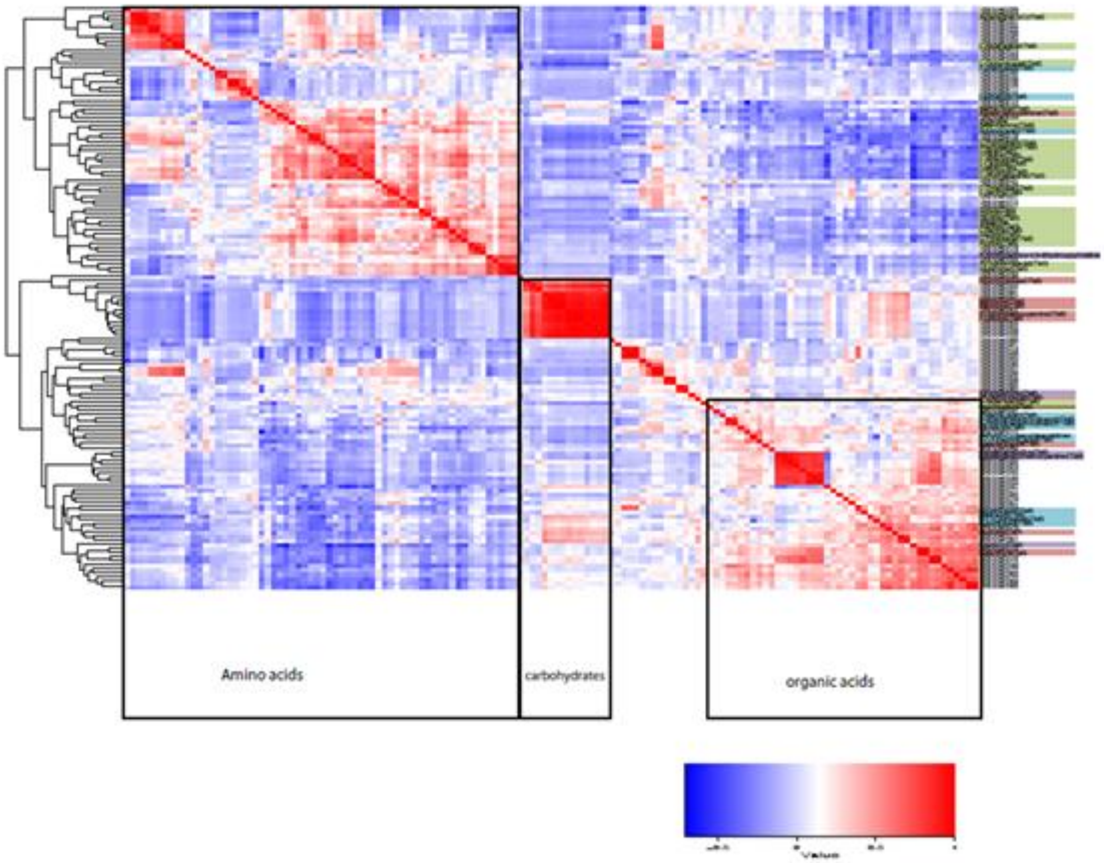


Figure 2: Correlation and cluster analysis of detected polar metabolites. Pearson correlation between metabolites is indicated by colour intensity,58 out of the 139 metabolites could be identified based upon spectral libraries and retention indices. Metabolites belonging to the same biochemical class (colour coded) tend to cluster together

C-parent map: 45 significant mQTLs were detected in this map for 39 masses representing unique metabolites. For 33 metabolites, only one QTL was identified and a maximum of two QTLs were found for six different metabolites (identified as galactiric acid and five unknown metabolites (No. 012, 034, 044, 078 and 081). The largest number of mQTLs on a single chromosome was 11 for chromosome eight, where mQTLs were detected for asparagine, tyrosine and other unidentified metabolites. The mQTL for tyrosine on chromosome eight was also detected in a previous study on the same diploid population (Werij et al., 2007). On chromosome five, we found eight mQTLs for glutamic acid, mannose, tryptophan and a number of unknown compounds. On chromosomes three, six and ten no significant mQTLs were detected. Fig. 3a shows the QTL profiles of all the metabolites mapped to the C parent map in a heat map.

Table 1 Metabolites putatively identified from the polar phase of methanol extract from potato tubers.List of metabolites that were putatively identified based upon similarity of mass spectra and the retention index publish in literature. Compounds that were detected in more than one derivitized form are listed only once.

Compound class	Metabolites
Amino acid	Alanine, asparagine, glutamine, glycine, aspartic acid, glutamic acid, isoleucine, leucine, methionine, phenylalanine, proline, serine, threonine, tryptophan, tyrosine, valine, lysine, pyroglutamic acid, β-Alanine, γ-Aminobutyric acid (GABA)
Organic acid	2-Piperidinecarboxylic acid (pipecolic acid), ascorbic acid, butanoic acid, citric acid, quinic acid, fumaric acid, glyceric acid, malic acid, phosphoric acid, succinic acid, dehydro-L-(+)-ascorbic acid dimer, lactic acid, threonic acid
Sugar	Allose, fructose, galactiric acid, galactinol, mannose, sucrose, glucopyranose
Sugar alcohol	Myo-inositol
Amino alcohol	Ethanolamine
Other	Calystegine B2, 5-Aminocarboxy-4,6-dihydroxypyrimidine, Allantoin

E-parent map: In this map, 160 significant mQTLs were detected for 85 representative masses (Fig. 3b). For 33 masses, only one QTL was detected and a maximum of six mQTLs for one metabolite (identified as quinic acid). The highest number of mQTLs on a single chromosome was 71 on chromosome five. This chromosome also contributed the most to the total explained variation of all detected mQTLs. A single genomic region, spanning three adjacent markers, accounted for the highest density of detected mQTLs (34). This region is known to be involved in plant maturity (Van Eck and Jacobsen, 1996; Collins et al., 1999; Oberhagemann et al., 1999) and as such exerts many pleiotropic effects on developmental related traits. The majority of compounds mapping to this region were classified as amino acids, organic acids and carbohydrates. This is not unexpected, as rapid changes in primary metabolites are known to occur during the later stages of maturation. Interestingly, similar to observations in the C-map, some amino acids that are biochemically related shared an mQTL at the plant maturity locus, e.g., glycine and threonine. Other amino acids like phenylalanine, lysine, valine and methionine also mapped to the plant maturity region. Some of the identified compounds mapping to this region also showed significant mQTLs in other chromosomes, both in the C and the E map. For example, four more mQTLs were detected for L-threonine, in chromosomes one, two and ten in the E-parent map and chromosome seven in the C-parent map.

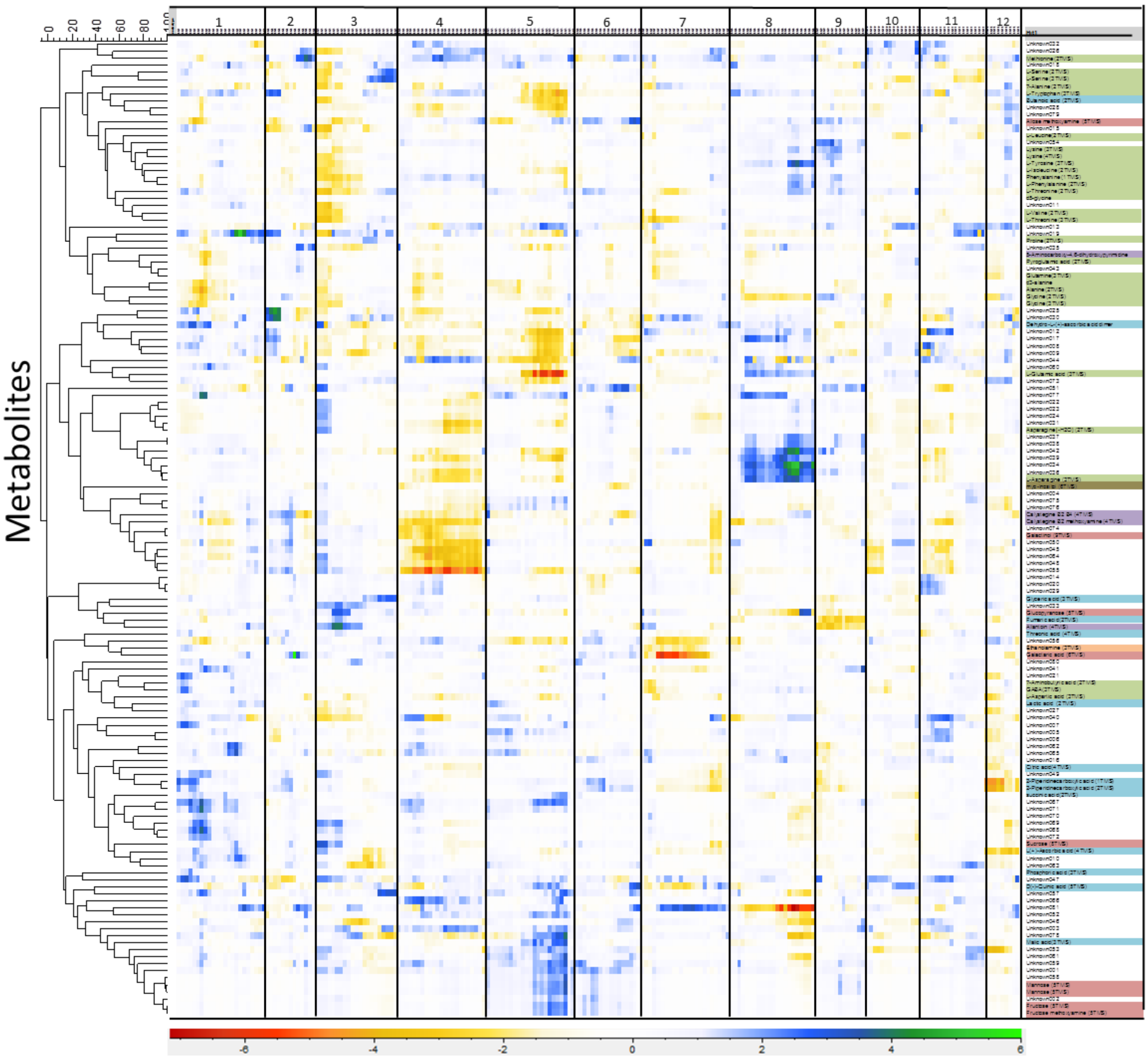
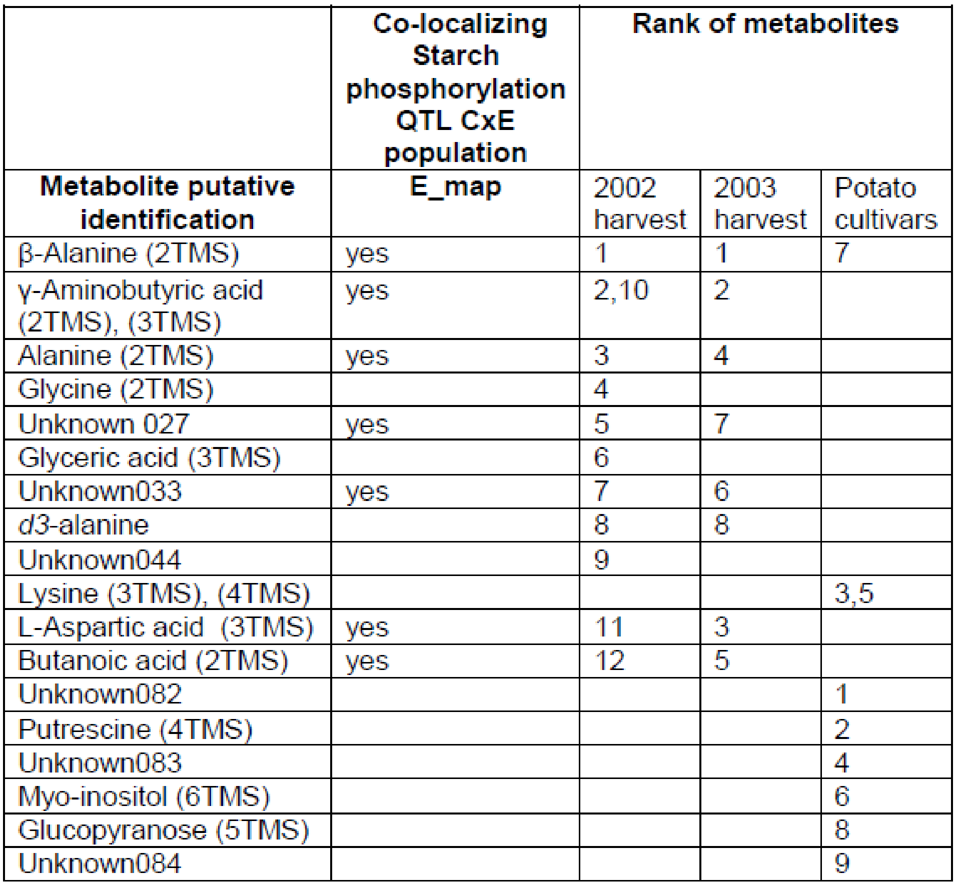


Figure 3a: Heat map of mQTL profiles of the detected compounds in the C x E population using the genetic map of the C-parent. Metabolites are clustered according to -10log(p) values across significantly associated markers. Vertical lines indicate chromosomal borders. Colours indicate the sign of the additive effect and the significance of the mQTL.

For methionine another mQTL was detected in chromosome two of the C-parent; and for valine one additional mQTL was detected in chromosome three of the E-parent indicating complex regulation of these traits.

Table 2 List of associated metabolites ranked after Randon Forest and significance permutation tests.



	Co-localizing Starch phosphorylation QTL CxE population	Rank of metabolites		
Metabolite putative identification	**E_map**	2002 harvest	2003 harvest	Potato cultivars
β-Alanine (2TMS)	yes	1	1	7
γ-Aminobutyric acid (2TMS), (3TMS)	yes	2,10	2	
Alanine (2TMS)	yes	3	4	
Glycine (2TMS)		4		
Unknown 027	yes	5	7	
Glyceric acid (3TMS)		6		
Unknown033	yes	7	6	
d3-alanine		8	8	
Unknown044		9		
Lysine (3TMS), (4TMS)				3,5
L-Aspartic acid (3TMS)	yes	11	3	
Butanoic acid (2TMS)	yes	12	5	
Unknown082				1
Putrescine (4TMS)				2
Unknown083				4
Myo-inositol (6TMS)				6
Glucopyranose (5TMS)				8
Unknown084				9

Association between phenotypic and metabolic traits

Traits of agronomical importance in potato breeding for tuber quality such as starch and cold sweetening are expected to be associated with primary metabolites. To investigate this relationship in more detail we carried out a parallel QTL analysis for phenotypic starch and cold sweetening related traits determined for this population for two years of harvest: 2002 and 2003 (Supplemental Table S4). Having mapped both metabolic and phenotypic QTLs to the two parental maps, we investigated the level of co-regulation of these two sets of traits by determining co-localizing QTLs. As expected, a substantial number of the phenotypic traits mapped to the plant maturity locus at chromosome five. However, a number of significant QTLs for essential metabolites were also detected outside this region indicating a possible regulation independent of the developmental stage and therefore these mQTLs are of particular interest.

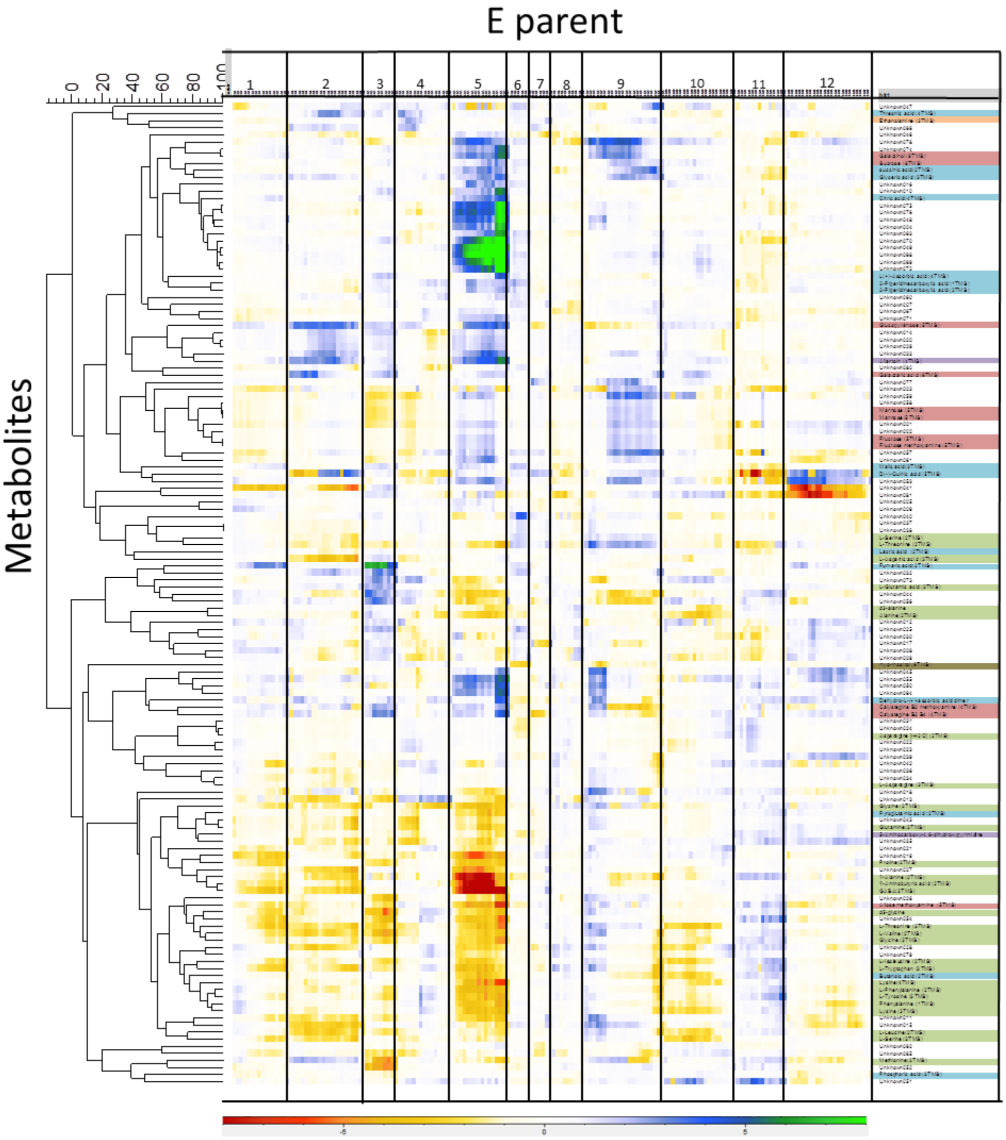


Figure 3b**:** Heat map of mQTL profiles of the detected compounds in the C x E population using the genetic map of the E-parent. Metabolites are clustered according to -10log(p) values across significantly associated markers. Vertical lines indicate chromosomal borders. Colours indicate the sign of the additive effect and the significance of the mQTL.

The evaluation of the co-localizations revealed several interesting shared QTLs between metabolic and phenotypic traits. From a total of 26 phenotypic traits, nine showed QTLs co-localizing with mQTLs outside the plant maturity region on chromosome five (Supplemental Table S5). Five phenotypic traits mapped to the C-map of which three mapped to the same position on chromosome eight: specific gravity of starch (also mapping to chromosome ten in the E-parent map) and decolouration of tuber flesh after cooking at 5 and 30 minutes. This position coincided with mQTLs for two unknown compounds (No 044 and 081). A fourth trait, starch grain particle size, mapped to a different region of chromosome eight on the C-map which co-localized with mQTLs for amino acids, organic acids and carbohydrates. Chip colour difference between reconditioning and harvest mapped to chromosome six on the C parent map and chromosomes three and nine on the E-parent map but only the latter two regions co-localized with a limited number of mQTLs.

Two traits, chip colour after harvest and chip colour difference storage-harvest mapped to chromosomes nine and three of the E-parent map. These positions coincide with genomic regions where mQTLs for succinic acid, fumaric acid, butanoic acid and Unknown044 were detected.

The strongest association, however, was detected between starch phosphorylation and a number of metabolites. starch phosphorylation maps to positions on chromosome two and nine of the E-parent map in both harvest years 2002 and 2003. The detection of identical QTLs in independent experiments suggests a strong and robust genetic control, although a third QTL on chromosome five was detected in the 2003 harvest which was only suggestive in 2002. This QTL co-localizes with mQTLs for alanine, butanoic acid, β-alanine, GABA (γ-amino butyric acid), succinic acid, pyroglutamic acid, phenylalanine, glutamine, tyrosine, tryptophan and seven unknown compounds (No 011, 027, 033, 048, 061, 068 and 069). The QTL for phosphate content on chromosome two co-localizes with mQTLs for serine, threonine, aspartic acid, GABA (γ-amino butyric acid), glutamine and four unknown metabolites (No 015, 027, 041 and 043). The QTL on chromosome nine co-localizes with a mQTL for galactinol.

Random Forest analysis reveals a link between primary metabolites and starch phosphorylation

To further investigate the strength of the co-localizations we focused on the phenotypic trait for starch phosphorylation (i.e. the degree of phosphorylation per mg of starch). Potato starch is characterized by a relatively high content of phosphate groups comparing to, for example, cereal starches (Rooke et al., 1949) (Hizukuri et al., 1970). This level of phosphate groups influences the viscosity and the chemical properties of starch and therefore it is important for the different uses of potato starch for food and industrial applications. Starch phosphorylation was determined in two subsequent years (2002, 2003). A strong correlation (R^2 = 0.8) between both years was found indicating a general high reproducibility of the expression of this trait. The observation that two to three QTLs could be mapped in each year together with a high broad sense heritability (H^2 = 0.5) further indicates that a substantial part of the trait variation can be attributed to genetic factors. In addition, a number of co-localizing mQTLs were identified (see above) that suggest links between starch phosphorylation and metabolic processes. To rank the associations between starch phosphorylation and the 139 representative masses, we used a multivariate Random Forest (RF) regression approach (Breiman 2001). The starch phosphorylation measurements of the two years were used separately as a response variable regressed against the 139 representative masses over all population individuals and significantly associated metabolites were recorded .

Using starch phosphorylation measurements of the 2002 harvest and all of the detected compounds by GC-TOF, the RF model explained 16 % of the variance (prediction R^2) at a permutation threshold of α=0.001. Twelve metabolites were significantly associated with phosphate content at this threshold (Table 2). Univariate correlation analyses between these significant metabolites and starch phosphorylation yielded absolute R^2 values ranging between 0.07 and 0.26. (Of which two have a negative Pearson correlation value). For the 2003 harvest the RF model explained 33 % of the variance at a permutation threshold of α=0.001. In this case, eight metabolites were found to be significantly associated with starch phosphorylation (Table 2). The correlations here ranged from 0.12 to 0.39 (of which only one has a negative Pearson correlation value). Interestingly, all the significantly contributing metabolites from the 2003 model were also identified using the 2002 data, again illustrating the high reproducibility between the two years. From these eight compounds seven showed co-localizing QTLs with at least one of the starch phosphorylation QTLs: β-alanine, GABA (γ-aminobutyric acid), L-aspartic acid, alanine, butanoic acid, Unknown027 and Unknown033 (Supplemental Table S5).

As a third independent line of investigation we performed RF regression on a subset of cultivars from a potato collection available in our laboratory (year of harvest 2007). 214 of this collection of cultivars was analysed for starch phosphorylation and 30 cultivars were selected covering the whole distribution range of this trait.(Supplemental Fig. S1) This selected set was analyzed for metabolite content, using the same analytical procedure as applied for the C x E population. The RF regression was performed using the starch phosphorylation measurements and the representative masses identified for this set of cultivars (data not shown). The resulting RF model explained 30% (prediction R^2) of the variance in starch phosphorylation and nine compounds were found to contribute significantly at a permutation threshold of α=0.001 (Table 2). The univariate correlations between these significant compounds and starch phosphorylation ranged between 0.09 and 0.41 (of which four have a negative Pearson correlation value) .

A comparison of the significant predictive compounds after RF analysis in the two sets of potato material (i.e. C x E (2002, 2003) and cultivar collection) revealed one compound in common. This compound was identified as β-alanine. In the C x E population β-alanine shows a positive correlation with starch phosphorylation in both years. This positive trend was also observed in the selected cultivar set (Supplemental Fig. S2). Because a robust metabolic predictor of a phenotypic feature is preferably valid across different potato sources we consider β-alanine as a reliable metabolite significantly linked to the level of phosphorylation of starch in potato tubers.

Discussion

The use of an untargeted metabolomics approach permits a quantitative assessment of a wide range of metabolites and allows the detection of unknown metabolites involved in known biochemical pathways. Untargeted metabolomics approaches have been successfully applied to experimental plant populations to uncover loci controlling variation of metabolites. (Overy et al., 2005; Keurentjes et al., 2006; Morreel et al., 2006; Schauer et al., 2006; Tieman et al., 2006; Lisec et al., 2008; Rowe et al., 2008)

In this study we used untargeted GC-TOF-MS metabolite profiling to assess the quantitative variation of polar metabolites present in dormant tubers of a diploid potato population. The observed variation in this population enabled us to locate genomic regions involved in the regulation of a range of polar primary metabolites. Primary metabolites, consisting mainly of carbohydrates, amino acids and organic acids, have an essential role in plant growth and development. In potato, carbohydrates are important for a number of agronomic traits related to tuber quality such as starch content and cold sweetening. In this study, the same genetic material was used to detect QTLs for starch and cold sweetening related traits and metabolic traits.

We investigated associations between phenotypic and metabolic traits through QTL co-localization, correlation and Random Forest analyses. The detection of a QTL identifies the existence of at least one polymorphic locus that is contributing to the variation observed for a given trait (Causse et al., 2004). When QTLs for two different traits co-localize this might indicate the existence of a common regulator that controls the variation of both traits. This is of special value in the search of candidate genes for traits with complex modes of inheritance or for which most of the genetic basis is unknown. However, it cannot be excluded that co-localizing genomic regions contain genes that are closely linked but are involved in different biological processes. Due to the limited resolution of QTL mapping and a finite number of markers, co-localization of QTLs is inevitable when large data sets are involved (Lisec et al., 2008). Therefore we performed independent statistical tests to confirm true positive associations between metabolites and phenotypes. In addition, we have validated putative metabolic biomarkers in an independent set of potato varieties. Our stringent selection criteria resulted in the determination of a strong relationship between potato starch phosphorylation and primary metabolism. Furthermore, our analyses resulted in identification of β-alanine as an important predictor for the degree of phosphorylation of starch in potato tubers.

Mapping of metabolic variation in potato tubers

Mapping populations are very suitable to identify loci controlling the variation of a given trait. In this study we aimed to explore the variation of metabolic and phenotypic traits present in dormant potato tubers. Our results show that we could assign genomic regions involved in metabolite variation for approximately 72% of the detected metabolites.

The abundances of metabolites that share an mQTL, especially for major loci of qualitative traits, are expected to correlate because they co-segregate in a mapping population. For instance, L-leucine and lysine share an mQTL on chromosome nine and are positively correlated. Metabolites sharing an mQTL often belong to the same biochemical pathway. For example, phenylalanine and tyrosine share a common biosynthetic pathway and hence are found to be co-regulated. Alternatively, co-locating QTLs can be the result of closely linked independent regulators (Lisec et al., 2008). In case of a shared regulator a direct, or even causal, relationship can be expected between traits whereas in the latter case the two traits are independently controlled. This distinction should be reflected in correlation analyses with higher values expected for co-regulated traits. In contrast, high correlation values between traits can also be expected when environmental factors that affect multiple traits simultaneously are in play. The resulting decrease in signal to noise ratio would be reflected in low heritabilities and QTL detection power. We therefore have applied independent lines of investigation, including genetic, correlation and Random Forest analyses to reliably identify biologically meaningful relationships between metabolites and complex phenotypes.

In targeted studies, QTLs were found for some of the metabolites that were also detected in our analyses. In a previous study an mQTL for tyrosine was detected on chromosome eight in the C parent map (Werij et al., 2007). This amino acid has been reported to be associated with levels of enzymatic discoloration (Corsini et al., 1992), although other studies have reported otherwise (Mondy and Munshi 1993). In agreement with the results of Werij et al. 2007 we did not find overlapping QTLs for tyrosine and enzymatic discoloration. In addition, we confirmed the QTL detected by Werij et al. 2007 and also detected two more QTLs at chromosomes five and eleven of the E-parental map. This difference is likely to be explained by differences in analytical techniques used in the two studies. Additionally, revisions in the linkage map that was used in our study could have influenced the power of detection of QTLs.

Another interesting metabolite that was also mapped in previous studies is methionine. The levels of this amino acid are related to the nutritional value of potato tubers. Moreover, it is the precursor of metabolites important to potato flavour (e.g., methional) and attempts have been made to enhance the methionine content for this purpose (Di et al., 2003; Dancs et al., 2008). In earlier work on the C x E population two QTLs were detected underlying the variation of this amino acid content in tubers (Kloosterman et al., 2010). The QTLs detected in that study mapped on chromosomes three and five, which is in agreement with our findings.

The significant mQTLs detected in both parental maps were unequally distributed over the genome, indicating hot and cold spots for metabolite regulation. A well-known locus involved in plant maturity is located at chromosome five, where a major QTL for this trait has been detected in the C x E population (van Eck and Jacobsen 1996). Plant maturity has been shown to be closely linked to a number of traits, including resistance and developmental traits (Collins et al., 1999; Oberhagemann et al., 1999; Bormann et al., 2004), although the underlying genetic factor has not been identified thus far. Products of primary metabolism such as carbohydrates and amino acids are expected to influence the degree of plant maturity and vice versa. We therefore anticipated that a large number of metabolic traits would show association with the plant maturity region on chromosome five. Nonetheless, a substantial number of mQTLs for amino acids, organic acids and carbohydrates have not been reported before and were identified outside this region. This finding highlights the importance of other genomic regions in the regulation of primary metabolite accumulation despite the pleiotropic effects displayed by the plant maturity region.

In addition, a number of mQTLs mapped to multiple positions, which indicates complex regulation. Among the multiple loci detected for these metabolites, at least one mapped to the plant maturity region. This raises the question whether metabolites are under developmental control or whether development is under metabolic control. The fact that many metabolites map at the maturity locus in addition to multiple other loci and plant maturity only at one may indicate that metabolism is at least partly under developmental control or possibly another factor upstream.

Putative predictors of starch phosphorylation

QTL co-localizations can be useful to identify metabolites involved in the regulation of phenotypic traits. This is of special importance for traits of which the genetic basis is unknown providing a valuable tool to search for candidate genes. However, one should be cautious when making such assumptions because two different traits that share the same regulatory region are not necessarily involved in the same molecular or biological process. In a specific genomic region genes might be present that are linked but that have different enzymatic functions. The phenotypic traits evaluated in this study are known to be related to carbohydrate metabolism and consequently metabolites involved in this pathway are likely to be linked to these traits. Nevertheless, QTL co-localizations can disclose unknown associations and moreover identify candidate predictors of trait variation (Lisec et al., 2008). One of the aims of this study was to test to what extent phenotypic and metabolic QTLs co-localize in order to identify metabolites truly associated to phenotypic features. We focused on the measurements for starch phosphorylation as a phenotypic case study. Potato starch has a particularly high content of phosphate groups in comparison to other plant species. The degradation of starch is dependent on reversible phosphorylation of the glucans at the surface of the starch granule (Zeeman et al., 2010) and although a direct link between the content of phosphate groups and starch degradation has not been found it has been shown that alterations in the starch phosphorylating enzymes lead to an excess of starch accumulation in the plant (Caspar et al., 1991; Zeeman and Rees 1999; Yu et al., 2001). In potato the high phosphate content of starch affects the viscosity and formation of stable starch pastes (Wiesenborn et al., 1994; Viksø-Nielsen et al., 2001) which is important for the diversified uses of starch in industry.

Here, we show that a number of amino acid mQTLs co-localise with trait QTLs for starch phosphorylation. To measure the strength of the genetically inferred links between the detected metabolic and phenotypic QTL co-localizations we examined the associations and predictive power of the metabolite data for starch phosphorylation using RF regression analysis.

The application of multivariate statistical methods to assess associations between metabolites and phenotypic traits has been successfully applied in a number of studies. An approach using canonical correlation analysis to test the predictive power of metabolic composition for biomass traits in Arabidopsis revealed a number of metabolites related to biomass and growth (Meyer et al., 2007). In potato a Partial Least Squares (PLS) analysis was used to discover metabolites that function as predictors for susceptibility to black spot bruising and chip quality (Steinfath et al., 2010).The validity of these results was tested in a collection of potato cultivars and in a set of individuals of a segregating population where metabolic and phenotypic information obtained from a first environment was used to predict phenotypic properties from the metabolic data obtained from a second environment. Those results demonstrate the application of multivariate data analysis and the value of independent validation to discover a small set of metabolites that can be used as biomarkers for a phenotypic trait of interest.

We used Random Forest analyses to predict starch phosphorylation from a GC-MS data set A similar approach was used to predict flesh colour and enzymatic discoloration from transcriptomics and liquid chromatography – mass spectrometry (LC-MS) data set (Acharjee et al., 2011). This study resulted in the successful identification of associated genes and metabolites, of which some were known to be involved in the regulation of the traits under study. Correspondingly, in our study, the application of RF regression resulted in a list of highly ranked metabolites, representing the most important compounds associated with starch phosphorylation. Inspection of the annotation of these included a number of unknown metabolites and more interestingly a few amino acids for which we also detected mQTLs coinciding with phosphate content QTLs. Amongst these relevant metabolites, β-alanine was of particular interest because it consistently ranked in the top metabolites in the different potato materials used for the analysis.

Starch phosphorylation is mainly driven by the action of two glucan-,water dikinases (i.e. Glucan water dikinase; GWD and Phosphoglucan water dikinase; PWD). These enzymes are critical in the transfer of phosphate groups within amylopectin (Smith et al., 2005; Zeeman et al., 2010). Analyses of Arabidopsis mutants also showed that GWD is required for phosphorylation (Yu et al., 2001). The *sex1* (loss of GWD activity) and *pwd* (loss of PWD activity) mutants lead to excess and reduced starch content phenotypes, respectively. Interestingly, transgenic potato plants displaying a reduced expression level of GWD also showed a starch excess phenotype (Lorberth et al., 1998). In our results the amino acid β-alanine was highly ranked after RF analysis for both sets of potato material (i.e. the C x E segregating population (2002 and 2003) and the set of potato cultivars). What is more, an mQTL for β-alanine was detected in the C x E population co-localizing with a phenotypic QTL for starch phosphorylation measurements of 2003 and a suggestive QTL in 2002. Although no specific role of either of these amino acids in starch phosphorylation has been reported, it is known that GWD follows a dikinase type reaction catalysing the transfer of the *β*-phosphate of ATP to either the C6 or C3 position of the glucosyl residue (Ritte et al., 2002). In this type of reaction the formation of an autophosphorylated intermediate precedes the transfer of the phosphate to the glucosyl residues. The autophosphorylation of this GWD intermediate depends on a conserved histidine residue which, when replaced by alanine results in a mutant phenotype without phosphorylating activity (Mikkelsen et al., 2004). Alanine isomers were further suggested as phosphate carriers when reacting with cyclo-triphosphate (P3m) to form orthophosphate derivates in high pH conditions (Tsuhako et al., 1985). These studies suggest a role for alanine in phosphorylation reactions although further research is needed to confirm these relationships. β-alanine, as a substrate for pantothenate (vitamine B_5) biosynthesis, is the only naturally occurring β-amino acid in plants (Chakauya et al., 2006). In plants, little is known about the formation of β-alanine, while in bacteria β-alanine is synthesized from the decarboxylation of L-aspartate in a reaction catalysed by aspartate decarboxylase (ADC) (Chakauya et al., 2006). Interestingly, we observed a shared mQTL for β-alanine and L-aspartate, suggesting common genetic regulation through shared biosynthesis pathways.

After a dormant phase, potatoes develop from a sink to a source organ that will subsequently support the growth and development of the new sprout. Owing to a higher content of phosphate groups, starch may be more easily mobilized and converted into resources for the growing sprout. Vitamine B_5 is used in the synthesis of coenzyme A (CoA), an acyl carrier protein. CoA is required in many central metabolic processes and it is essential in the conversion of pyruvate to acetyl-CoA to enter the tricarboxylic acid cycle (TCA cycle) (Chakauya et al., 2006). In addition, CoA is fundamental in the biosynthesis of fatty acids, polyketides, depsipeptides and peptides (Kleinkauf 2000). β-alanine constitutes an important part in the biosynthesis pathway of Vitamine B_5 and the presence of this amino acid may be indicative for the formation of many essential metabolites for plant development and furthermore, it may act as an indicator of the mobilization of storage resources. In this study, we identified β-alanine associated with phosphate content as well as a number of other metabolites for which it also might be predictive. Our approach has been shown to be instrumental in generating hypotheses about functional relationships between metabolites and phenotypes. In addition it may help for a gradual understanding of metabolic processes contributing to observed phenotypical features of interest.

Our current data demonstrate the benefits of the applied methods for a broad untargeted metabolomics approach in potato. In this study we combined genetic information through mQTL and phenotypic QTL analysis and non-genetic information through regression of trait values to predict phenotypic traits from metabolomics analysis. We identified candidate metabolites which can be informative for phenotypic traits of interest.

The advances in metabolomics have opened up the way to high-throughput approaches allowing the analysis of variation of a large number of samples in a reasonable amount of time. In addition, advanced statistical methods enable us to explore and monitor different profiling techniques in non-model species. Furthermore, the genome sequence of potato (Xu et al., 2011) has now revealed genes specific to this highly heterozygous crop, bringing a platform that will ultimately facilitate the elucidation of the genetic basis of complex traits of high importance in breeding for tuber quality.

Supplemental Data

The following materials are available in the online version (http://www.plantphysiol.org/content/158/3/1306/suppl/DC1) of this article

Supplemental Table S1. Identification of detected polar primary metabolites
Supplemental Table S2. List of detected metabolic QTL detected in C maternal linkage map
Supplemental Table S3. List of detected metabolic QTL detected in paternal linkage map
Supplemental Table S4. List of phenotypic QTLs for starch and cold sweetening related traits.
Supplemental Table S5. Phenotypic and metabolic QTL co-localizations
Supplemental Table 6 List of starch and cold sweetening related traits measured in the C x E population.
Supplemental Figure S1. Plots of response ratios of phosphate content to different levels of β-alanine.
Supplemental Figure S2. Frequency distribution of phosphate content for potato cultivars.

Acknowledgements

We gratefully acknowledge the technical and intellectual input of our laboratory colleagues, Francel Verstappen, Anna Undas, Desalegn Etalo, Benyamin Houshyani and Hassanzadeh. Special thanks to Aaron Velez-Ramirez for his help in the preparation of the figures. We also thank Thierry Delatte for his valuable comments and suggestions on the discussion.

Chapter 5

Genetical genomics of quality related traits in potato tubers using proteomics

Animesh Acharjee[1,2], Luc Suurs[2], Pierre-Yves Chibon[1,2], Bjorn Kloosterman[2,6], Twan America[3,4], Jenny Renaut[5], Chris Maliepaard[2], Richard G.F. Visser[2,3]

[1]Graduate School Experimental Plant Sciences

[2]Wageningen UR Plant Breeding, Wageningen University and Research Center, PO Box 386, 6700 AJ Wageningen, The Netherlands

[3]Centre for BioSystems Genomics, P.O. Box 98, 6700 AA Wageningen, The Netherlands,

[4]Plant Research International, Wageningen University and Research Center, P.O. Box 16, 6700 AA Wageningen, The Netherlands

[5]Centre de Recherche Public - Gabriel Lippmann Department of Environment and Agrobiotechnologies (EVA) 41, rue du Brill, L-4422 Belvaux, Luxembourg

[6]Current address: Keygene NV, PO Box 216, 6700 AE Wageningen, The Netherlands

To be submitted

Abstract

Recent advances in ~omics technologies such as transcriptomics, metabolomics and proteomics along with genotypic profiling have permitted the genetic dissection of complex traits such as quality traits in non-model species. To get more insight into the genetic factors underlying variation in quality traits related to carbohydrate and starch metabolism and cold sweetening, we determined the protein content and composition in potato tubers using 2D-gel electrophoresis in a diploid potato mapping population. We performed pQTL analyses for all proteins with a sufficient representation in the population and established a relationship between proteins and twenty-six potato tuber quality traits (*e.g.* flesh colour, enzymatic discoloration) by co-localization on the genetic map and a direct correlation study of protein abundances and phenotypic traits. We were able to map pQTLs for over 300 different protein spots some of which co-localized with traits such as starch content and cold sweetening. pQTLs were observed on every chromosome. For some 20 protein spots multiple QTLs were observed. Hotspot areas for protein QTLs were identified on chromosomes three, five, eight and nine. The hotspot on chromosome 3 coincided with a QTL previously identified for total protein content and had more than 23 pQTLs in the region from 70 to 80 cM. Some of the co-localizing protein spots associated with some of the most interesting tuber quality traits were identified.

Key words: Genetical genomics, proteomics, protein QTL, potato quality traits

1 Introduction

Potato *(Solanum tuberosum* L.*)* is the third most important food crop consumed worldwide. It is vegetatively propagated by means of tubers which develop from underground stems called stolons that under favourable conditions enlarge and increase in size and shape to form tubers. The active growth and development of tubers is accompanied by important changes in the physiology and genetic regulation that lead to large depositions of starch and storage proteins (Prat et al., 1990; Visser et al., 1994). The nutritional and industrial value of the tubers is mainly from their carbohydrate content which comprises 80% starch along with nutritionally important concentrations of essential amino acids and Vitamin C (Scott et al., 2000). Considering the large amount of storage proteins of the tubers, a proteomics approach was chosen as a suitable way to study potato for specific tuber quality traits, additional to studies already performed using LC-MS and GC-MS profiling of the tubers (Acharjee et al., 2011; Carreno Quintero et al., 2012).

Quantitative trait locus analysis has been applied to levels of gene expression enabling the identification of genomic loci controlling the observed variation in gene expression (eQTLs). This approach was called ‘genetical genomics’ (Jansen and Nap 2001; Doerge 2002; Schadt et al., 2003). Similar approaches can be followed for data derived from other ‘~omics’ technologies such as proteomics (resulting in pQTLs, protein QTLs) and metabolomics (mQTLs, metabolite QTLs) (Keurentjes et al., 2006; Kliebenstein 2007).

We generated proteomics data from a well-studied diploid potato mapping population (here denoted as C x E) using 2D-DIGE (two-dimensional gel electrophoresis). We mapped the variation in protein levels by treating these levels as quantitative traits in a QTL analysis.

In addition, we performed a QTL analysis for several quality related traits (including starch and cold sweetening), to study co-location of protein QTLs and phenotypic QTLs. These are traits for which in many cases there was no prior knowledge with respect to which genes might regulate or determine these traits. Identifying metabolites or proteins may then help in getting an idea about the potential genes involved. We identified pQTL and phenotypic QTL (phQTL) hotspot areas (Breitling et al., 2008) across the potato genome and detected pQTLs that co-localized with phenotypic QTLs. Through identification of the proteins and combining the protein QTL (pQTL) results with QTLs from phenotypic traits (phQTL) we can acquire knowledge about the genes and/or proteins which are controlling the variation in quantitative phenotypic traits. In addition, we study the direct correlation between the phenotypic traits and the protein intensities. This approach offers a tool for plant breeders to get insight into the genetics of complex traits which primarily depend on protein content, constitution, and/or expression. We made a first attempt for the identification of some of these co-localizing protein spots.

2 Materials and Methods

2.1 Plant material

A diploid potato (*Solanum tuberosum* L.) population C (USW5337.3) X E (77.2102.37) was used consisting of 98 individuals (detailed explanation can be found in the materials and methods sections of chapters 2, 3 and 4). The genotypes were grown in the field in 2002 and 2003 and the tubers were harvested. All clones were grown in Wageningen, The Netherlands during the normal potato growing season (April–September). For each genotype, all tubers were collected from three plants and representative samples were either used for phenotypic analyses or mechanically peeled and immediately frozen in liquid nitrogen before being ground into a fine powder and stored at -80°C for subsequent proteomic analysis.

2.2 Phenotypic analyses

Different quality traits are considered in the phQTL study. A detailed list of phenotypic traits that were assessed can be found in the supplementary Table 1. In this study, we focus on 26 quality traits related to starch characteristics (11 traits) and colour and cold sweetening (15 traits). A detailed description of how the different traits were assessed and analyzed can be found in Celis-Gamboa 2005, Werij 2011 and Werij et al., 2011.

2.3 Proteomics data generation and processing

2.3.1 Protein extraction

Total protein was extracted from approximately 0.5 gram of ground tuber material, to which 1 ml of pre-heated (95°C) lysis buffer (50mM sodium phosphate buffer pH 7, sucrose (5% w/v), SDS (4% w/v), DTT (0.3% w/v), PVP-P (10% w/v)) was added. Samples were homogenized for 45 seconds, placed at 95°C in water bath for 1 minute and homogenized again (45 seconds, speed 6.5 m/sec). After 3 minutes at 95°C in water bath the samples were cooled on ice and centrifuged for 15 minutes . 4 ml acetone (-20°C) containing 10mM DTT was added to the supernatant, vortexed vigorously and put at -20°C for 1 hour. The protein extract was centrifuged for 20 minutes in a Centricon T42-k (25000g, 4°C). The pellet was washed with 4 ml acetone -20°C containing 10mM DTT twice. After air drying the pellet, the pellet was dissolved in 300 µl TUCCDT buffer (urea 5M, thio-urea 2M, C7BzO (2% w/v), CHAPS (2% w/v), DTT (0.3% w/v), TCEP 2mM). Protein amount was measured using the RC/DC assay (Biorad, Veenendaal, the Netherlands) using Bovine SerumAlbumine (BSA) as standard for the calibration curve.

2.3.2 Protein labelling

A single lysine per protein molecule was labelled using the fluorescent CyDyes from the Difference Gel Electrophoresis (DIGE) technology (GE Healthcare/Amersham Biosciences) according to the manufacturer's protocol. The internal standard was labelled with Cy2 and consists of an equal mixture of 20 randomly chosen samples of the experiment (9 random samples from 2002 and 2003 each and both parents C and E from 2003).

Every 2D-gel contains one sample labelled with Cy3, one labelled with Cy5 and the internal standard labelled with Cy2. This means that every sample on each gel can be compared by using the internal standard sample labelled with Cy2.

2.3.3 2D-Electrophoresis

The first dimension electrophoresis was performed using 24cm immobilized pH gradient strips (GE Healthcare/Amersham Biosciences) with a linear pH range from 4 to 7 on an Ettan IPGPhor isoelectric focusing (IEF) system. Cydye labelled samples (total of 150µg protein) were loaded to the strips diluted in 0.5% IPG buffer (pH4-7 and pH3-10, 1:1) and TUCCDT buffer to a volume of 450µl. The focusing was run for 18 hrs at 20°C with the following settings: 3 h 150V, 3 h 300V, from 300V to 1000V in 6 h, from 1000V to 10000V in 1 h and finally 5 h at 10000V. After IEF the strips were equilibrated in the dark at room temperature in equilibration buffer (urea 6M, 50mM Tris-HCl pH8.8 , glycerol 30% (v/v), SDS 2% (w/v)) containing DTT 1% (w/v) for 15 minutes and after that in the same buffer with added containing iodoacetamide 2.5% (w/v) for 15 minutes. The second dimension electrophoresis was run on the Ettan Dalt twelve system on precast 12.5% SDS polyacrylamide slab gel (size: 255x196x1 mm) and buffers from GE Healthcare/Amersham Biosciences. Electrophoresis was performed at 1W/gel for 1 h followed by 1.5W/gel until bromo phenol blue had reached the end of the gel (approximately 17 h) at 15°C. The separated CyDye-labelled proteins were visualized by scanning with a Ettan Dige Imager (GE Healthcare/Amersham Biosciences), using for Cy2 an 480 nm laser and an emission filter of 530nm, for Cy3 an 540 nm laser and an emission filter of 595 nm and for Cy5 an 635 nm laser and an emission filter of 680 nm.

2.3.4 Image analysis and data pre-processing

Gel images were analyzed with the Decyder software version 7 according to Decyder 2Dv7 manual; GE Healthcare/Amersham Biosciences. The detected spots were then filtered based on spot volume larger than relative value 30000 to exclude spots that could be just background noise or dust particles. The internal standard in each gel was used to automatically match all images to the reference (the gel with the largest number of detected spots). After that a gel area with saturated spots coming from patatin was excluded because this protein was at the ceiling level of detection for all samples as this is a very abundant storage protein. To make 2D-spot alignment across the samples a clear image gel was chosen as the master and added to all the gel batches (1 batch is one run of 12 gels). Then these batches were linked to each other by automatic matching in the software program and correcting afterwards manually with the help of setting landmarks (i.e. spots visible in all images). The spot volume ratio to the internal standard of each protein and the individual volume of the spots were calculated and $\log_{10}$ transformed. In the QTL analysis the spot volume (intensity) value was used. Each of the proteins are presented by Pro_X where "X" represents consecutive protein numbers, numbered from top to bottom and from left to right starting with number 1 in the top left and ending with number 1643 in the right bottom of the gel.

2.3.5 Protein identification

Spots of interest were excised from gel using the Ettan Spot Picker. After washing and desalting in 50 mM ammonium bicarbonate/50% v/v methanol, followed by 75% v/v ACN, spots were digested with Trypsin Gold (MS grade, Promega, Madison, WI, USA, 8 mg.mL-1 in 20 mM ammonium bicarbonate) using the Ettan Digester robot. Automated spotting of the samples was carried out with the spotter of the Ettan Spot Handling Workstation Peptides dissolved in a 50% ACN containing 0.5% TFA (0.7 mL) were spotted on MALDI-TOF disposable target plates (4800, ABSciex, Foster City, CA, USA) prior to the deposit of 0.7 mL of CHCA (7 mg/mL, 50% v/v ACN, 0.1% v/v TFA, Sigma Aldrich, St. Louis, MO, USA). Peptide mass determinations were carried out using the Applied Biosystems 4800 Proteomics Analyzer. Both PMF and MS/MS in reflectron mode analyses were carried out with the samples. Calibration was carried out with a peptide mass calibration kit.. Proteins were identified by searching against the NCBI 'viridiplantae' database (15334873 sequences, September 2011) and an EST 'viridiplantae-eudicots' database (75859188 sequences, October 2010) using MASCOT. All searches were carried out using a mass window of 50 ppm for MS and 0.75 Da for MS/MS. The search parameters allowed for carboxyamidomethylation of cysteine as fixed modification, and oxidation of methionine as variable modification. Homology identification was retained with a probability threshold of 95%, all identifications were manually checked.

2.5 QTL mapping

QTL mapping of the protein spots was done based on the spot volume ratio to the internal standard (intensity) of the proteins (after transforming the different spots into a quantitative value). QTL analysis was done using the R/qtl library (Broman and Sen 2009). A genome-wide LOD significance threshold (4.28) was computed using the Li & Ji algorithm (2005) and was used for all QTL analyses. The data was loaded in R and run through the jittermap function from R/qtl and probabilities of the underlying genotypes were computed using a hidden Markov model, as available in the calc.genoprob function of R/qtl with a step size of 2.5 cM. We performed the “4way” (terminology used in R/qtl for a cross between two heterozygous diploid parents) procedure for simple interval mapping using the Haley-Knott regression method (Haley and Knott 1992). Significant QTLs (LOD > 4.28) were extracted and the explained variances of these QTLs were computed. For each QTL the following information was reported: start position (cM position where the significance threshold was passed), peak cM position, and stop position (cM position where the LOD score drops under the significance threshold again), start, peak and stop marker, LOD value for the peak marker and the explained variance (R^2) at the peak position. More detailed information is provided by Kloosterman et al., 2012.

The genetic map used in the QTL analyses consisted of 343 markers. This is a modified version of an earlier C x E genetic map (Anithakumari et al., 2010), with all sequence based SNP markers and extended with additional markers from allele specific hybridization signals using a potato microarray (Anithakumari et al., 2010). In order to describe the density of pQTLs and phQTLs over the genome, we calculated numbers of pQTLs or phQTLs using a 10 cM sliding window according to Chen et al., 2010.

We considered pQTLs to be co-localized with phQTLs if they fell within a 10 cM interval (5 cM to the left and 5 cM to the right) around the peak marker of the phQTL.

3 Results

In this study, we generated proteomics data from 2D-DIGE (Difference gel electrophoresis). The patatin protein family (storage proteins in the potato tuber (Liu et al., 2003) was left out for further analysis because of the overabundance of these proteins, clearly visible as a large block of different proteins in the middle of the gel (Fig. 1A). Initially 1643 unique spots were detected in total over the two harvests of 2002 and 2003. We considered the two year harvests to see the consistency and/or difference in the pQTLs. We did pQTL analysis with 380 protein spots for the 2002 harvest and 320 spots for the 2003 harvest that were measured in all samples. The number of overlapping spots for both years was 255 with an extra 125 observed only for 2002 and an extra 65 unique to 2003. The highest absolute Pearson correlation coefficient among these 255 proteins was 0.98 for both 2002 (between protein nr. 39 and 40) and 2003 (between protein nr. 295 and 297). The analysis was done as described in the materials and methods section by taking a quantitative measure of the different spots (log10 of spot volume) and analyze them for the individual 90 genotypes as can be seen from the example in Figure 1B.

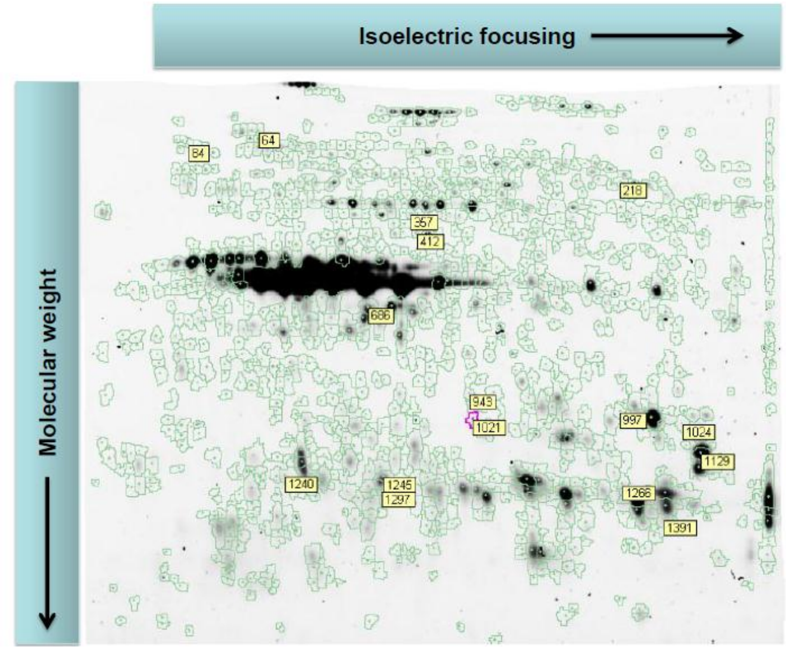


Figure 1A: Example of a 2D DIGE gel image. Different protein spots which are co-localizing with a flesh colour QTL are shown in yellow boxes. The dark protein spots in the middle and left of the gel are the over-abundant patatin proteins.

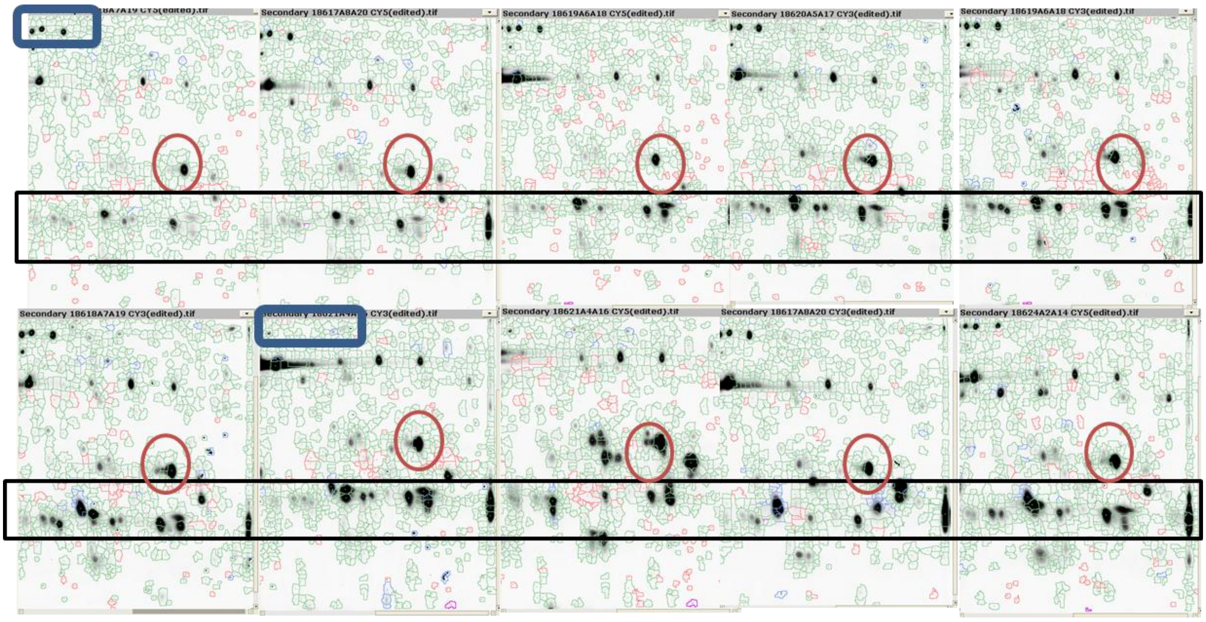

Secondary 18617A8A20 CY5(edited).tif
Secondary 18619A6A18 CY5(edited).tif
Secondary 18620A5A17 CY3(edited).tif
Secondary 18619A6A18 CY3(edited).tif
Secondary 18618A7A19 CY3(edited).tif
CY3(edited).tif
Secondary 18621A4A16 CY5(edited).tif
Secondary 18617A8A20 CY3(edited).tif
Secondary 18624A2A14 CY5(edited).tif

Figure 1B: A panel of 10 different gels (in all cases the right bottom quarter is depicted in this figure) showing the patterns and the absence/presence (blue rectangular area), always present in almost similar amounts (round area), as well as several examples of spots with varying quantities over different samples (black rectangular area).

For 2002, 190 significant pQTLs were found for 170 protein spots (113 spots from the 255 common ones, 57 from the protein spots unique for 2002). For the 2003 harvest, we found 173 pQTLs for 154 protein spots (130 from the 255 common ones, 24 spots that were unique for 2003). We found 82 pQTLs that mapped in the same chromosome in both years and out of these 82, 56 pQTLs mapped in the exact same position (identical peak position) in the chromosome across two years (supplementary Table 2).

We found 20 proteins for the 2002 harvest with QTLs on two different chromosomes. For 2003, 17 proteins had QTLs in either two or three different chromosomes and those proteins gave 36 pQTLs in total. Out of these 17 proteins, 2 had 3 different pQTLs and 15 proteins had 2 pQTLs each.

Comparing pQTLs from the 2002 and 2003 harvests separately, for 2002 the largest percentage variation explained (R^2) for a pQTL was 94%, for protein number 429, and this pQTL is mapped to chromosome 7 at 7.4 cM. For 2003, this pQTL maps to the same position with 70% explained variance. The Pearson correlation coefficient in the abundance of this protein between the two years was 0.81. The QTL with the largest amount of variance explained for the 2003 harvest (74%) was for protein number 1007 and this QTL mapped to chromosome 3. For 2002, this pQTL was found in the same chromosome and same position explaining 88% of the variance. The Pearson correlation in the abundance of this protein between the two years was 0.75.

In both years the largest number of pQTLs was found for chromosome 8 (41 and 31 pQTLs, for 2002 and 2003, respectively) and the lowest for chromosome number 10 (2 and 3 pQTLs, for 2002 and 2003 respectively).

To investigate if the pQTLs were evenly distributed across the genome, or clustered in particular regions, we calculated the density of pQTL per cM across the genome using a 10 cM sliding window analysis (Fig. 2A). For the 2002 harvest, four regions had a high pQTL density centering around markers PotSNP749 (position: bch (Chr. 3, 80 cM), PotSNP125 (Chr. 5, 23 cM), PotSNP749 (Chr. 8, 6 cM) and STM3012 (Chr. 9, 16 cM), each having more than 8 pQTL per cM which is much higher than the expected 0.17 pQTLs per cM if the 190 pQTLs were evenly distributed along the 1135 cM genetic linkage map.

For the 2003 harvest a total of 152 pQTLs (88% of the 173 significant pQTLs) are mapped on chromosomes 3, 5, 6, 7, 8, 9, 11 and 12 and the number of pQTLs were 30, 24, 11, 11, 31, 21, 11 and 13 respectively (Fig. 2B). Similar to the 2002 harvest, we observe that hotspot regions of pQTLs are found for chromosome numbers 3, 5, 8 and 9.

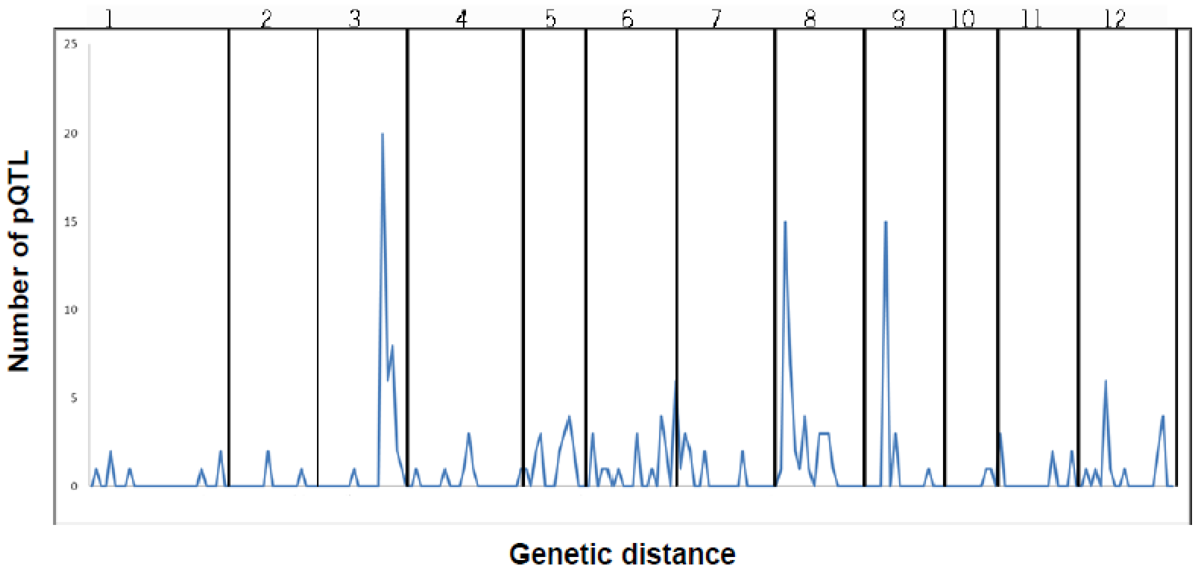


Figure: 2A

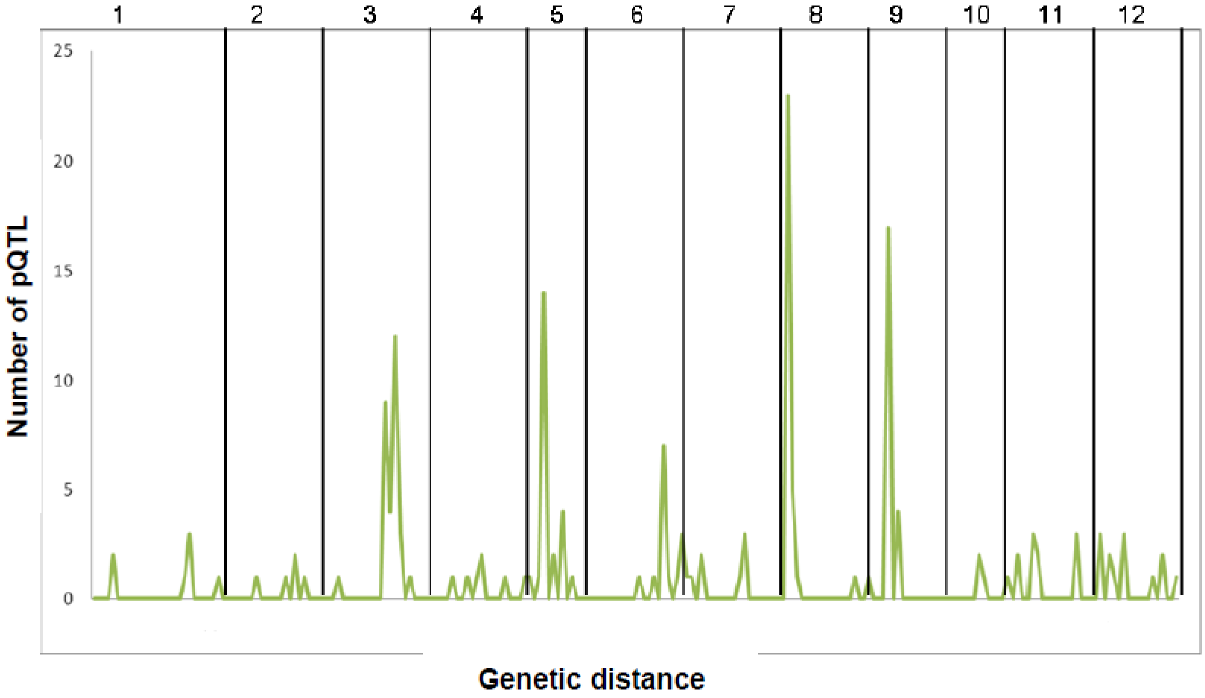


Figure: 2B

Figure 2A and 2B: Protein QTL (pQTL) density for proteomics data (2002 in panel A; 2003 in panel B) across the genome. x-axis: pQTL genomic location on chromosomes. y-axis: pQTL density calculated on a 10 cM sliding window. Chromosomal regions corresponding to the largest number of significant pQTLs are considered pQTL hotspots, on chromosomes 3, 5, 8, 9

3.1 Correlations of protein spots with quality traits:

From the Pearson correlation study among protein spots and quality traits 22 protein spots were significantly correlated to quality traits with FDR corrected (Benjamini and Hochberg 1995) p value ($p < 0.05$ for the FDR corrected t-test on Pearson correlation). In total 10 protein spots were found significantly correlated with flesh colour of which the highest correlation coefficient was 0.67 for protein number 1129 and the lowest significant correlation coefficient 0.34 for protein number 686. The highest correlation coefficient with enzymatic discoloration after 30 min and 3 hours both were 0.44 for protein spot number 1129. Four protein spots were significant for enzymatic discoloration after 30 min and 3 hours. Four protein spots were significant with for starch phosphorylation. The highest correlation coefficient for starch phosphorylation was with protein number 129 (r=0.44).

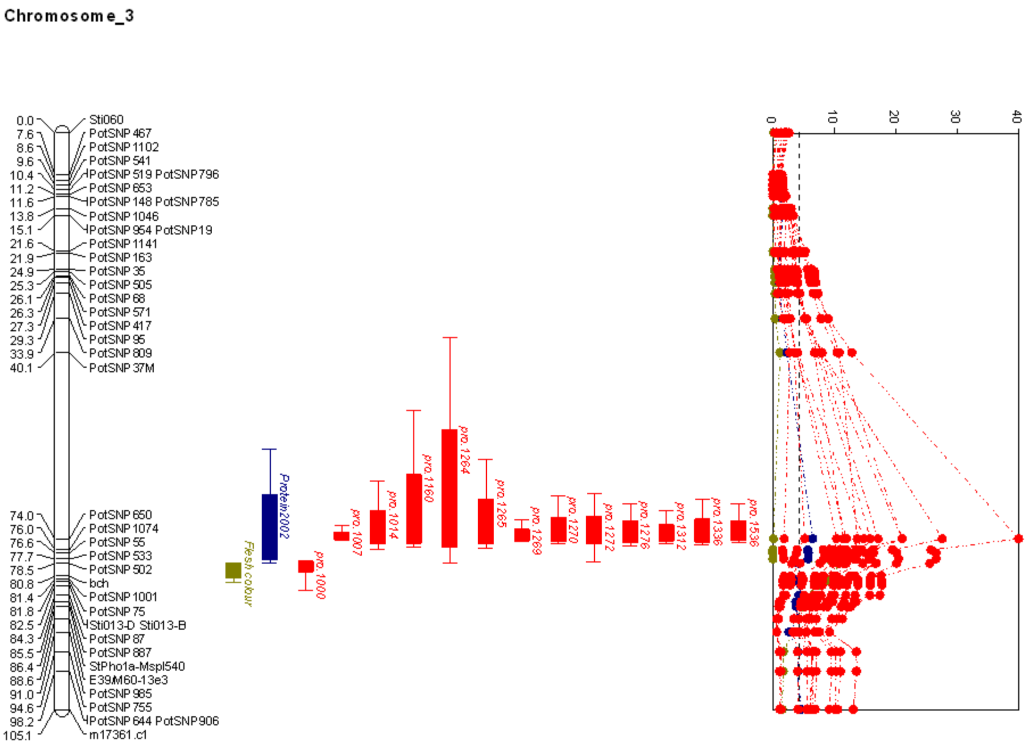
3.2 Phenotypic QTL (phQTL) analysis:

QTLs for the majority of the starch related quality traits such as percentage of amylose and starch gelatinization related traits are mapped to chromosome 2, specifically in the region between 73.7 cM and 80.2 cM (start and end position). A single QTL for flesh colour and enzymatic discoloration is mapped to chromosome 3, in the region between 78.5 cM and 81.4 cM (Brown et al., 2006). We did not find any significant phQTLs for the quality traits studied here on chromosomes 4, 7, 9, 11 and 12. Detailed results of the QTL analyses for starch and cold sweetening related traits are presented in supplementary Table 3.

We focused on co-localizations of phQTLs related to starch traits, (enzymatic) discoloration and cold sweetening and pQTLs for the analyses of two years (2002 and 2003). Such co-localizations can be useful to identify proteins involved in the regulation of these phenotypic traits. One other striking observation was that a phenotypic QTL for total protein content (Werij et al., 2012) on chromosome 3 in the region of 70-80 cM corresponded approximately with 23 different pQTLs (one of the four hot-spot regions of pQTLs).

For the 2002 harvest: in chromosome 1, a QTL for starch gravity is co-localized with two proteins (pro_375 and pro_102) between 126.6 cM and 135.0 cM. QTLs for percentage of amylose and starch gelatinization related traits co-localize with a pQTL on chromosome 2 in the region between 73.7 and 80.2 cM. QTLs for flesh colour and enzymatic discoloration (after 5 and 30 minutes) are co-localized with 14 pQTLs on chromosome 3 between 78.5 and 88.5 cM (Fig. 3A and 3B). On chromosome 5, phenotypic QTLs for differential scanning calorimetry and chip colour after harvest are co-localized with two pQTLs at 23.6 cM. A QTL for starch-phosphorylation is also co-localized on chromosome 5 with 9 other pQTLs between 40.3 cM to 54.8 cM. A QTL for particle size distribution of the starch is co-localized on chromosome 6, between 56.4 cM to 59.9 cM with 3 pQTLs. A QTL for specific gravity of starch is co-localized with a pQTL on chromosome 8, in the region of 59.2 cM to 67.8 cM. For the 2003 harvest: in chromosome 1, a QTL for starch gravity is co-localized with a QTL for Protein 1240 between map positions 126.6 and 135.0 cM. QTLs for the percentage of amylose and starch gelatinization related traits are co-localized with QTLs for three protein spots on chromosome 2 between 73.7 and 80.2 cM. QTLs for flesh colour and enzymatic discoloration (after 5 and 30 minutes) co-localize with QTLs for 31 protein spots on chromosome 3 between 74.0 to 88.5 cM. On chromosome 5, QTLs for differential scanning calorimetry and chip colour after harvest co-localize with QTLs for 15 protein spots in the region between 40.3 and 54.8 cM. A QTL for starch-phosphorylation is also co-localized in chromosome 5 with QTLs of five other protein spots in the region between 40.3 and 51.5 cM. A QTL for particle size distribution of starch is co-localized in chromosome 6, at exactly the same position (56.4 cM) with a QTL for Protein 251 for both the years. We did not find co-localization of any protein QTLs with the QTL for specific gravity of starch for the 2003 harvest. Detailed results are shown in supplementary Table 4 for the 2002 harvest and in supplementary Table 5 for the 2003 harvest.

A



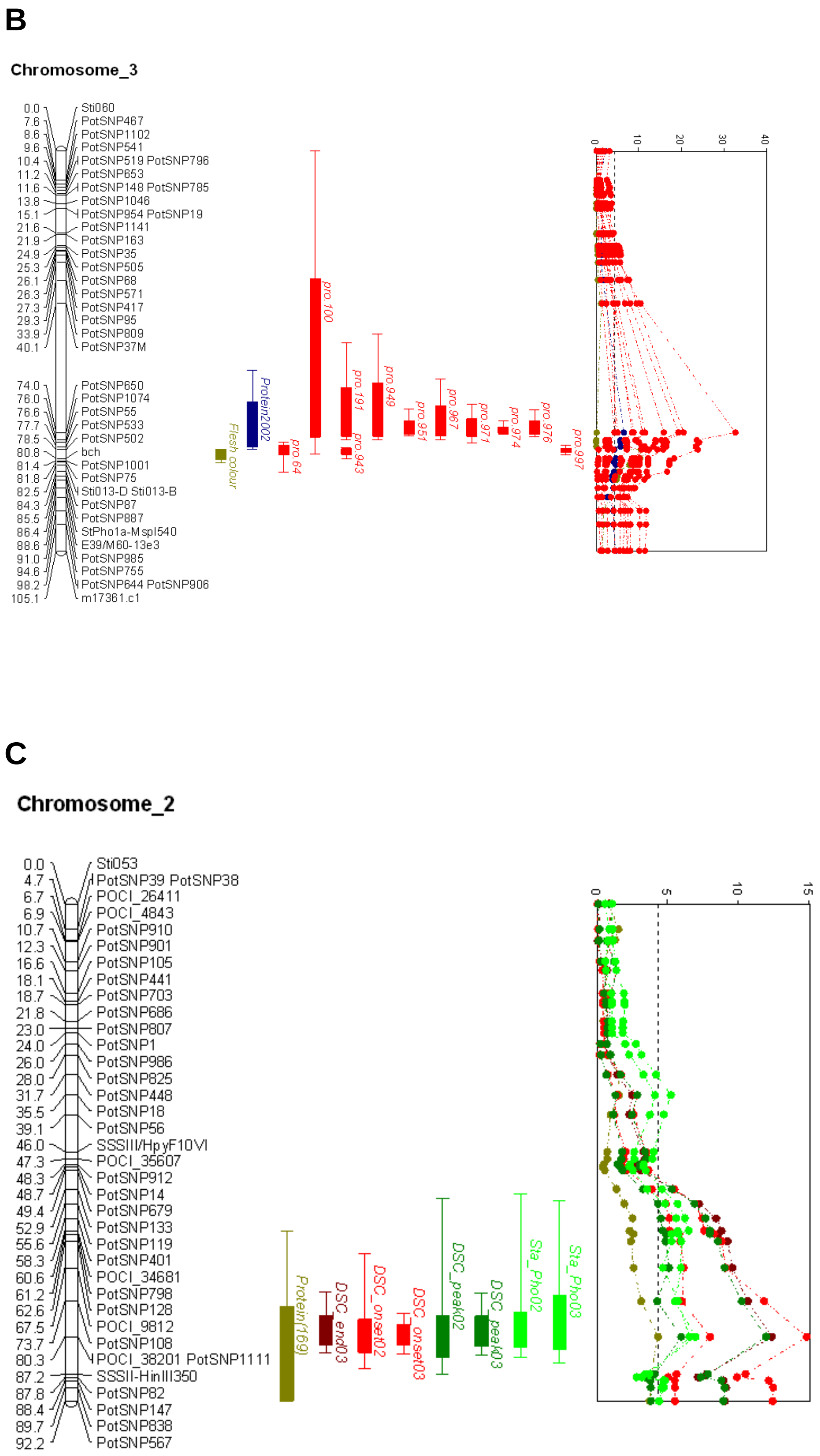

B
Chromosome_3
0.0 Sti060
7.6 PotSNP467
8.6 PotSNP1102
9.6 PotSNP541
10.4 PotSNP519 PotSNP796
11.2 PotSNP653
11.6 PotSNP148 PotSNP785
13.8 PotSNP1046
15.1 PotSNP954 PotSNP19
21.6 PotSNP1141
21.9 PotSNP163
24.9 PotSNP35
25.3 PotSNP505
26.1 PotSNP68
26.3 PotSNP571
27.3 PotSNP417
29.3 PotSNP95
33.9 PotSNP809
40.1 PotSNP37M
74.0 PotSNP650
76.0 PotSNP1074
76.6 PotSNP55
77.7 PotSNP533
78.5 PotSNP502
80.8 bch
81.4 PotSNP1001
81.8 PotSNP75
82.5 Sti013-D Sti013-B
84.3 PotSNP87
85.5 PotSNP887
86.4 StPho1a-MspI540
88.6 E39/M60-13e3
91.0 PotSNP985
94.6 PotSNP755
98.2 PotSNP644 PotSNP906
105.1 m17361.c1
Flesh colour
Protein2002
pro.64
pro.100
pro.191
pro.943
pro.949
pro.951
pro.967
pro.971
pro.974
pro.976
pro.997
0
10
20
30
40
C
Chromosome_2
0.0 Sti053
4.7 PotSNP39 PotSNP38
6.7 POCI_26411
6.9 POCI_4843
10.7 PotSNP910
12.3 PotSNP901
16.6 PotSNP105
18.1 PotSNP441
18.7 PotSNP703
21.8 PotSNP686
23.0 PotSNP807
24.0 PotSNP1
26.0 PotSNP986
28.0 PotSNP825
31.7 PotSNP448
35.5 PotSNP18
39.1 PotSNP56
46.0 SSSIII/HpyF10VI
47.3 POCI_35607
48.3 PotSNP912
48.7 PotSNP14
49.4 PotSNP679
52.9 PotSNP133
55.6 PotSNP119
58.3 PotSNP401
60.6 POCI_34681
61.2 PotSNP798
62.6 PotSNP128
67.5 POCI_9812
73.7 PotSNP108
80.3 POCI_38201 PotSNP1111
87.2 SSSII-HinIII350
87.8 PotSNP82
88.4 PotSNP147
89.7 PotSNP838
92.2 PotSNP567
Protein(169)
DSC_end03
DSC_onset02
DSC_onset03
DSC_peak02
DSC_peak03
Sta_Pho02
Sta_Pho03
0
5
10
15

Figure 3 A to C: A) Example of the abundance of pQTLs (QTLs are indicated with 2 LOD support intervals) on chromosome 3. pQTLs are shown in red, the QTL for total protein content in blue, for flesh colour in light green.

B) Continuation of example of the abundance of pQTLs on chromosome 3.

C) Another example of the abundance of pQTL on chromosome 2 is shown. QTLs for different quality traits such as differential calorimetry (DSC onset in red, DSC peak in dark green) and starch phosphorylation (in light green) are co-localized with protein number 169.

We tried to identify more proteins, especially those associated with enzymatic discoloration and flesh colour. These attempts were not very successful although we were able to get an amino acid sequence for some of the proteins. In most cases the putative identities of these proteins did not make immediate sense but in the case of enzymatic discoloration enzyme functions like chaperonin (protein nr 239), protein disulfide isomerase (nr 280), aminoaldehyde dehydrogenase (nr 200), plastidic phosphoglucomutase (nrs 171 & 175) and methionine synthase (nr 62) were retrieved which are among the types of functions which one could imagine that might be involved in this specific pathway. However more research into this area is required.

Table 1: List of proteins isolated from gel and putatively identified. Full name of the traits are listed in the supplementary Table 1.

Protein Nr.	Est	Protein name	Traits associated with the proteins or pQTLs from the proteins in the first column
39	NA	5-lipoxygenase [Solanum tuberosum]	Multiple pQTL 2003
62	NA	methionine synthase [Solanum tuberosum]	Enz. Discol5min
64	NA	NA	Flesh Colour
128	Gi|58217733	gi|108709562|gb|ABF97357.1| Lysyl-tRNA synthetase, putative, expressed [Oryza sativa Japonica	Multiple pQTL 2002
129	NA	RecName: Full=Transketolase, chloroplastic; Short=TK; Flags: Precursor	Starch_Phos03
175	gi|10808429	gi|8250622|emb|CAB93680.1| plastidic phosphoglucomutase [Solanum	Enz. Discol5min

		tuberosum]	
186	NA	NA	Multiple pQTL 2002
180	NA	NA	DSCdH03
200	NA	aminoaldehyde dehydrogenase 2 [Solanum lycopersicum]	Enz. Discol5min
196	NA	NA	Enz. Discol5min
193	NA	importin alpha, putative [Ricinus communis]	Multiple pQTL 2002
237	NA	chaperonin-60 beta subunit [Solanum tuberosum]	Enz. Discol5min
218	NA	NA	Flesh Colour
280	gi|14644452	gi|4704766|gb|AAD28260.1|AF131223_1 protein disulfide isomerase homolog [Datisca glomerata]	Enz. Discol5min
296	NA	vacuolar H+-ATPase B subunit [Nicotiana tabacum]	Top pQTL 2003
339	NA	ATP synthase F1 subunit 1 [Nicotiana tabacum]	Enz. Discol5min
372	NA	beta tubulin [Capsicum annuum]	Multiple pQTL 2002
379	NA	ATP synthase beta chain [Zea mays]	Starch.grT.2002
411	NA	NA	Enz. Discol5min
964	NA	catalase isozyme 1-like protein [Solanum tuberosum]	Multiple pQTL 2002
1021	gi|53697586	gi|161702915|gb|ABX76298.1| sexual organ expressed protein [Nicotiana alata]	Flesh Colour

4 Discussion

We did pQTL analysis with 380 proteins for 2002 and 320 proteins for the 2003 harvest separately and phQTL analysis for starch and cold sweetening related traits as well as flesh colour using an integrated linkage map of C x E. The pQTL analysis of the proteomics data resulted in a large number of genetic regions involved in protein abundance. The pQTLs are spread out over all chromosomes but four regions show a larger number of QTLs, so-called “hotspots” (Breitling et al., 2008). These hotspots contain most probably one causal factor for protein synthesis or regulation which maps to that locus (Chan et al., 2010b). In other plant species, for example in *Arabidopsis*, similar hotspots were detected after mapping transcripts, protein expression, metabolites, and phenotypic traits (Fu et al., 2009). These authors reported that the phenotypic variation was mainly due to six hotspots.

In our study, four hotspot regions consistent across the years 2002 and 2003 are found on Chr.3 near 70-80 cM, on Chr. 5, near 20-30 cM, on Chr. 8, position 6 cM and on Chr. 9, near 10-20 cM. This shows stability of pQTL hotspots across the two years. The fact that we find a hotspot for protein content as determined by Werij et al., 2012 with over 20 pQTLs may suggest that this concerns an overall regulator of protein synthesis in potato tubers. More research is needed to elucidate this.

In a previous study of expression QTLs (eQTLs) and metabolite QTLs (mQTLs) (Acharjee et al., 2011; Carreno Quintero et al., 2012) it was noted that the hotspot areas for expression and metabolites were mainly on chromosome 5 and 11. In the case of pQTL analysis we mainly find pQTL hotspots on chromosomes 3, 5, 8 and 9. This indicates that the genetic regulation of the protein expression and/or content are more likely controlled by specific locations on those chromosomes. Chromosome 5 is in common as a hotspot, for protein QTLs, metabolic and expression QTLs. Also for phenotypic QTLs including some of the agronomical traits, chromosome 5 is a hotspot (data not shown for agronomical traits: see Celis Gamboa 2005) due to pleiotropic effects of maturity or earliness on chromosome 5 (for pleiotropic QTLs for developmental traits see *e.g.* Hurtado 2012).

The phQTL on chromosome 3 for tuber flesh colour is consistent with earlier findings (Brown et al., 2006). Moreover, other reports link the gene beta-carotene hydroxylase with the QTL at this map position (Kloosterman et al., 2010; Wolters et al., 2010). Wolters et al., (2010) found one more gene involved in yellow tuber flesh colour: zeaxanthin epoxidase (Zep) on chromosome 2. They established this relationship in an association analysis between single nucleotide polymorphism (SNP) haplotypes and flesh colour phenotypes in a large range of diploid and tetraploid potato genotypes. In our analysis only half of the number of genotypes had tubers with yellow flesh colour and the statistical power may not have been enough for detecting this second QTL (Kloosterman et al., 2010).

In this study we used multi-year (2002 and 2003) protein and phenotypic datasets. For the 2002 harvest, 43% of the pQTLs are mapped to the same chromosome as the harvest of 2003, and, *vice versa*, 47% of the pQTLs from the 2003 harvest are mapped to the same chromosome as those for 2002. For 10 protein spots of the 2002 harvest we found 20 QTLs on two different chromosomes. Out of these 20 QTLs, ten have only a pQTL for the 2003 harvest; the other ten are mapped to the exact same position as for the 2002 harvest. For 2003, QTLs for each of 17 proteins mapped to either two or three different chromosomes and seven out of these proteins have a pQTL only for the 2002 harvest, the ten remaining ones were mapped to the exact same position as for 2003. The proteins which are mapped to the same chromosome across the years show consistency between the years. This is expected as the effects are genetic and the protein abundances are well-correlated across years. This indicates that genotype-by-environment interaction is not very large and that the measurement/technical variation is small in comparison with the genetic variation for these proteins.

4.1 Co-localization of pQTLs and phQTLs

In this study the same mapping population was used to detect phQTLs for carbohydrate related traits and protein traits. We investigated QTL co-localization between phenotypic and protein traits. As an example we have shown co-localization of a flesh colour QTL with QTLs of different protein spots (Fig. 3A and 3B). A detected QTL indicates a statistical association between a marker locus in that region and the quantitative variation for a given trait segregating in that same population (Causse 2004). When QTLs for two different traits co-localize, we could hypothesize the existence of a common locus that contributes to the variation of both traits, or we could consider the association to be due to linkage of different loci. Such hypotheses are useful in the search for candidate genes for phenotypic traits of interest for which most of the genetic basis is unknown.

In the correlation study, we found a protein, protein number 1129 to be positively correlated with flesh colour, and enzymatic discoloration after 30 min and 3 hours, with correlation coefficients 0.67, 0.44 and 0.44, respectively. A QTL for this protein was mapped to chromosome 3 at 80.8 cM for both years and co-localizes with the flesh colour QTL as well. From a previous study by Kloosterman et al., 2010, it was reported that carotenoids are involved in flesh colour and the beta-carotene hydroxylase (bch) gene plays a major role in flesh colour variation in potato. It is tempting to speculate that this protein would indeed be BCH but so far we were not able to identify this protein.

As already stated in the results section, identification of the proteins was not very successful. This was in part due to the necessity to run the gels a new and align them so that the spots could be picked. This proved to be more difficult than originally anticipated and thus more work (repeats) are required to elucidate the identity of more proteins.

In this paper we used genetic information from phQTL and pQTL analyses on the one hand and Pearson correlations of phenotypic traits with proteomics data on the other hand. From the QTL analyses, we can identify the map position of the QTLs but associations need not be from a functional relationship but can also be due to linkage. In correlating phenotypic traits to proteomics data, we find proteins that might be related to the phenotype, but in the absence of genetic information, this correlation could be due to environmental conditions influencing both the phenotypic trait and the protein abundance(s). Combining both QTL analysis and correlation analysis gives us a better understanding about candidate proteins which are linked to the phenotype but also show a genetic association. A similar type of approach was described in a study of Acharjee et al., 2011 where the authors combined QTL analysis with a prediction of the phenotypes from metabolomics and transcriptomics data using random forest regression.

Acknowledgements

We thank Dr. Carolina Blanca Celis Gamboa and Dr. Jeroen Werij for their contributions to phenotypic analysis. This work was financially supported by Wageningen UR Plant Breeding and the Centre for BioSystems Genomics through a proteomics ‘hotel’ grant.

Supplementary Table 1: List of carbohydrate and starch metabolism and cold sweetening related quality traits measured in the CxE population.

Colour and cold sweetening related quality traits	Description	Scoring
Flesh	Flesh colour	Flesh colour: 1= white cream; 2= cream;3= cream orange;4=yellow cream; 5=light yellow; 6=medium yellow; 7=dark yellow; 8=medium orange; 9=dark orange
Cc_4c	Chip colour after storage	Chip colour after tubers were stored for 3 months at 4 oC. 1= very dark brown to 10= very nice light golden colour
Cc_4c_rec	Chip colour after storage and reconditioning	Chip colour after cold storage at 4 oC during 3 months and then reconditioning for 3 weeks at room temperature (18-20C). 1= very dark brown to 10= very nice light golden colour
Cc_ah	Chip colour after harvest	Chip colour after harvesting: 1= very dark brown to 10= very nice light golden colour. From 7 colour acceptable by industry
Ccdif_4c_ah	Chip colour difference storage-harvest	Difference in chip colour between cold storage and harvest.

Ccdif_rec_4c	Chip colour difference recondition-storage	Difference in chip colour between reconditioning and storage at 4 oC.
Ccdif_rec_ah	Chip colour difference recondition-harvest	Difference in chip colour between reconditioning and harvest.
Discol_dif	Difference in discoloration	Difference in enzymatic discoloration between 3 hours and 5 minutes
Discol30min	Discoloration of flesh at 30 min	Enzymatic discoloration of raw flesh, after 30 minutes: 0= no change in flesh colour until 6=very dark red/brown colour
Discol3h	Discoloration flesh at 3h	Enzymatic discoloration raw flesh after 3h From 0= no change in flesh colour until 8= black flesh colour
Discol5min	Discoloration of flesh at 5 min	Enzymatic discoloration of raw flesh, after 5 minutes: 0= no change in flesh colour until 6=very dark red/brown colour
Discolcook_diff	Difference in discoloration after cooking	Difference in discoloration after cooking between 24 and 5 minutes values
Discolcook24h	Discoloration after cooking-24h	Same as Discolcook5m
Discolcook30min	Discoloration after cooking-30min	Same as Discolcook5m
Discolcook5m	Discoloration after cooking-5min	Discoloration after cooking, after 5 minutes: 1= no change in colour; 2= some light grey spots; 3= some darker grey spots; 4= light grey colour evenly distributed on the tuber; 5= dark grey colour evenly distributed on the tuber; 6=very dark grey/black colour evenly

		distributed on the tuber

Carbohydrate and starch metabolism related quality traits	Description	Scoring
DSC_T_diff	Differential Scanning Calorimetry	Starch gelatinization properties.
DSC_T_end	Differential Scanning Calorimetry	Starch gelatinization properties. End temperature of gelatinization
DSC_T_onset	Differential Scanning Calorimetry	Starch gelatinization properties. Onset temperature of gelatinization
DSC_T_peak	Differential Scanning Calorimetry	Starch gelatinization properties. Peak temperature of gelatinization
DSCdH	Starch gelling delta H	Temperature difference
PSD_d50	Particle_size_distribution midpoint	median of the particle size distribution of starch particles obtained from tubers
PSD_d90_d10	Particle_size_distribution d90-d10	Starch grain particle size distribution: difference between 90 percentile point and 10 percentile point (indicates distribution range)

Spec_grav_starch	Underwater weight derived dry matter content	Specific gravity: (5000xdry weight)*underwater weight. For chips, between 400-450; for French fries, between 450 and 500 and for starch industry >500
% Amylose	% amylose	Amylose fraction of starch, in percentage

Supplementary Table 2: pQTL analysis summary for 2002 and 2003 harvest

Year of harvest	Nr. of sig. pQTLs	Highest nr. of pQTL position	Lowest nr. of pQTL position	Nr. of multiple pQTLs
2002	190	Chrom 8 (41*)	Chrom 10 (2*)	20
2003	173	Chrom 8 (31*)	Chrom 10 (3*)	17

* Indicated numbers of pQTLs in that chromosome

Supplementary Table 3: Summary of Phenotypic QTL results of the various quality traits, their peak position and explained variance at peak position

Traits*	Year	Chr.	Peak LOD	QTL Peak Marker	QTL Peak Pos.	%Var(R2)
GemDS	2003	1	4.4	PotSNP1016	78.2	18.0
Starch_grT	2002	1	4.4	PotSNP558	126.7	19.3
Decol3h	1998	2	4.4	PotSNP133	52.9	21.6
DSC_T_end	2002	2	8.7	PotSNP128	62.6	37.1
%Amylose	2002	2	5.4	PotSNP108	73.7	25.1
%Amylose	2003	2	5.0	PotSNP108	73.7	23.8
DSC_T_onset	2002	2	8.0	POCI_38201	80.2	35.7
DSC_T_peak	2002	2	5.9	POCI_38201	80.2	28.3
Starch_Phos	2002	2	6.9	POCI_38201	80.2	31.8
DSC_T_onset	2003	2	14.8	POCI_38201	80.2	55.7
DSC_T_end	2003	2	12.4	POCI_38201	80.2	50.2
DSC_peak	2003	2	12.0	POCI_38201	80.2	48.7
Starch_Phos	2003	2	6.5	POCI_38201	80.2	30.1
Decol_diff	1998	3	4.5	PotSNP37M	40.1	18.7
Flesh	1998	3	10.5	PotSNP502	78.5	44.3
Decol5min	1998	3	5.5	PotSNP1001	81.4	26.4
Decol30min	1998	3	5.5	PotSNP1001	81.4	26.5
DSC_T_peak	2002	5	4.5	PotSNP125	23.6	21.2
Cc_4c	1998	5	5.7	PotSNP125	23.6	27.4
Starch_Phos	2003	5	6.6	PotSNP621	44.3	30.8
PSD_d9_d10	2002	6	5.0	PotSNP963	56.4	21.2
Spec_grav_starch	2002	8	4.4	PotSNP12	67.8	21.1
Spec_grav_starch	2002	10	5.1	PotSNP76	26.8	24.7

*Please find the explanation of these abbreviations and how the trait was scored in the supplementary table 1

Supplementary Table 4: Co-localization of phQTLs (starch, colour and cold sweetening related traits) and pQTLs for proteomics data in 2002

Traits	Chr. Nr.	QTL Peak Pos. (cM)
Starch_grT_2002	1	126.7
Pro_379	1	135.8
Pro_102	1	135.8
% Amylose_2003	2	73.7
%Amylose_2002	2	73.7
DSC_T_onset_2002	2	80.2
DSC_T_peak_2002	2	80.2
Starch_Phos_2002	2	80.2
DSC_T_onset_03	2	80.2
DSC_T_end_2003	2	80.2
DSC_T_peak_2003	2	80.2
Starch_Phos_2003	2	80.2
Pro_169	2	80.2
Discol3h	2	52.9
DSC_T_end_2002	2	62.6
Pro_736	2	49.3
Discol_diff	3	40.1
Pro_1282	3	40.1
Discol5min	3	81.4
Discol30min	3	81.4
Flesh colour	3	78.5
Pro_997	3	78.5
Pro_64	3	78.5
Pro_943	3	78.5
Pro_1000	3	78.5
Pro_1129	3	80.8
Pro_1245	3	80.8
Pro_1266	3	80.8
Pro_1391	3	80.8
Pro_1297	3	80.8
Pro_1240	3	80.8
Pro_1024	3	80.8
Pro_1021	3	80.8
Pro_6	3	85.5
Pro_152	3	85.5
DSC_T_peak_2002	5	23.6
Cc_4c	5	23.6
Pro_1045	5	23.6
Pro_1035	5	23.6
Starch_Phos_2003	5	44.3
Pro_129	5	40.3
Pro_128	5	40.3
Pro_142	5	48.2

Pro_848	5	48.2
Pro_143	5	48.9
Pro_144	5	53.3
Pro_140	5	54.8
Pro_150	5	54.8
Pro_928	5	54.8
PSD_d9_d10_2002	6	56.4
Pro_254	6	56.4
Pro_251	6	56.4
Pro_255	6	60.0
Spec_grav_starch	8	67.8
Pro_275	8	59.2

Supplementary Table 5: Co-localization of phQTLs (starch, colour and cold sweetening related traits) and pQTLs for proteomics data in 2003

Traits	Chr. Nr.	QTL Peak Pos. (cM)
Starch_grT_2002	1	126.7
Pro_1240	1	134.6
DSC_T_onset_2002	2	80.2
DSC_peak_2002	2	80.2
Starch_Phos_2002	2	80.2
DSC_T_onset_2003	2	80.2
DSC_T_end_2003	2	80.2
DSC-T_peak_2003	2	80.2
Starch_Phos_2003	2	80.2
% Amylose_2002	2	73.7
% Amylose_2003	2	73.7
Pro_152	2	73.7
Pro_188	2	73.7
Pro_272	2	80.2
Flesh colour	3	78.5
Discol5min	3	81.4
Discol30min	3	81.4
Pro_1007	3	74.0
Pro_1536	3	74.0
Pro_1269	3	74.0
Pro_1270	3	74.0
Pro_1160	3	74.0
Pro_971	3	74.0
Pro_951	3	74.0
Pro_491	3	74.0
Pro_1217	3	78.5
Pro_1416	3	78.5
Pro_1021	3	78.5

Pro_943	3	78.5
Pro_1217	3	78.5
Pro_1416	3	78.5
Pro_943	3	78.5
Pro_1129	3	80.8
Pro_1245	3	80.8
Pro_1091	3	80.8
Pro_1267	3	80.8
Pro_1000	3	80.8
Pro_1297	3	80.8
Pro_1240	3	80.8
Pro_1318	3	80.8
Pro_1391	3	80.8
Pro_1272	3	80.8
Pro_64	3	82.5
Pro_1317	3	82.5
Pro_152	3	86.4
Pro_153	3	86.4
Pro_1294	3	88.5
DSC_T_peak_2002	5	23.6
Cc_4c	5	23.6
Pro_1438	5	23.6
Pro_366	5	23.6
Pro_469	5	23.6
Pro_339	5	23.6
Pro_41	5	23.6
Pro_330	5	23.6
Pro_192	5	23.6
Pro_36	5	23.6
Pro_40	5	23.6
Pro_1051	5	23.6
Pro_39	5	23.6
Pro_357	5	23.6
Pro_666	5	20.1
Pro_1264	5	20.1
Pro_1317	5	18.4
Starch_Phos_2003	5	44.3
Pro_140	5	51.5
Pro_144	5	44.8
Pro_150	5	44.8
Pro_128	5	42.3
Pro_129	5	40.3
PSD_d9_d10_2002	6	56.4
Pro_251	6	56.4

Chapter 6

Prediction of potato quality traits from ~omics data

Animesh Acharjee[1,2], Bjorn Kloosterman[2,4], Richard G.F. Visser[2,3], Chris Maliepaard[2]

[1]Graduate School Experimental Plant Sciences

[2]Wageningen UR Plant Breeding, Wageningen University and Research Center, PO Box 386, 6700 AJ Wageningen, The Netherlands

[3]Centre for BioSystems Genomics, P.O. Box 98, 6700 AA Wageningen, The Netherlands,

[4]Current address: Keygene NV, PO Box 216, 6700 AE Wageningen, The Netherlands

To be submitted

Abstract:

In order to find genetic and metabolic pathways related to phenotypic traits of interest, we analyzed gene expression data, metabolite data sets obtained with GC-MS and LC-MS, proteomics data and a selected set of tuber quality phenotypic data from a diploid segregating mapping population of potato. In this study we present an approach to integrate these ~omics data sets for the purpose of predicting phenotypic traits. First we used Random Forest regression to select subsets of genes, metabolites and proteins showing a predictive association with the quality traits. Next, using a genetical genomics approach we identified eQTLs, mQTLs and pQTLs across the potato genome and established which of these co-segregate with phenotypic QTLs for the quality traits. Finally, in an integrated network analysis for each of the quality traits we selected representative sets of genes, metabolites and proteins from each of their eQTL, mQTL and pQTL regions and constructed correlation networks with these, including the quality trait. This gives us networks of relatively small sets of interrelated ~omics variables that can predict, with higher accuracy, a quality trait of interest. We validate this approach by comparing the selected genes, metabolites and proteins for the quality trait potato tuber flesh colour quality with the regulatory and metabolic pathways known to be involved in this trait.

Key words: genetical genomics, ~omics, random forest

1 Introduction

In order to understand how quantitative variation in phenotypic traits is related to the underlying genetic differences between plants, and to differences in gene expression, protein constitution and metabolic variability, an approach is needed in which the combined molecular signature of the plants is shown to be predictive for the phenotypic traits of interest (Fukushima et al., 2009; Kim et al., 2010). We can use high-throughput ~omics technologies, such as gene expression microarrays (Brazma and Vilo 2000; Gaasterland and Bekiranov 2000), mass spectrometry (LC-MS and GC-MS) (Fiehn 2002; Dunn et al., 2005) and protein chips (Aebersold and Mann 2003; Zhu et al., 2003) to obtain molecular signatures of a population of plants. In addition we can study phenotypic differences in the same population and hypothesize that differences in the phenotypic trait in the population are related to the variation in these combined molecular profiles across the different data sets. Finding the ~omics variables that are related to the phenotypic traits of interest can then be used in two ways: for prediction of the traits from these molecular profiles, and for identifying functional relationships between traits and molecular networks of the plants. This means that we are not just interested in interrelating these ~omics data sets for their own sake to find genetic and metabolic networks, but the networks and their elements should actually be predictive for a trait of interest. On the other hand, in the context of genetics and plant breeding, we also want to be protected against finding relationships between phenotype and ~omics data that are just caused by environmental or developmental differences. Instead, we are interested to find relationships that have a basis in the genetic differences between plants. Therefore a mapping population is an ideal target for this kind of study: we can study whether an observed relationship between a phenotypic trait and ~omics variables is also based on genetic differences between the plants in the segregating population, as we can actually map the phenotypic variation as well as the variation in the ~omics data sets. A relationship that would be just based on variation in environmental influences or conditions would not result in mapped QTLs for the phenotypic traits or the ~omics variables (Jansen and Nap 2001; Keurentjes et al., 2006).

We are interested in the association between a number of phenotypic traits related to tuber quality of potato and several ~omics data sets. In addition, we use mapping and genotyping information since the population that we use is a mapping population. The quality traits considered are 1) potato tuber flesh colour, 2) enzymatic discoloration after peeling, 3) starch gelatinization as measured by differential scanning calorimetry (DSC) and 4) tuber shape.

In this study, we try to relate transcriptomics, metabolomics and proteomics data to genetic variation in these quality traits in the mapping population. For this, we use a three-step strategy. First, using the same approach as in an earlier study (Acharjee et al., 2011), we apply Random Forest regression (RF) to find, per single trait and per individual ~omics data set, the variables that play a significant role in the prediction of each of the quality traits. Briefly, RF (Breiman 2001) is a collection of unpruned decision trees (Hastie et al., 2001) and can be used for statistical classification and regression. A RF model is typically made up of hundreds of decision trees and 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 or OOB samples). In the second step, we use QTL mapping of the quality traits and of the ~omics variables to select variables that have a QTL cosegregating with a quality trait QTL and to remove redundancy in the set of selected variables. Finally, we construct correlation networks, for each of the quality traits, of the selected sets of variables from the ~omics data sets that have a genetic association with the traits per QTL.

Because already much is known about the regulatory genetic and metabolic pathways involved in tuber flesh colour (Wolters et al., 2010; Werij et al., 2007; Kloosterman et al., 2012), we used this trait to validate the approach. We demonstrate that also for the other phenotypic traits we can find networks of small sets of interrelated gene expression profiles, proteins and metabolites that are associated with and predictive for these quality traits.

2 Materials and Methods

2.1 Plant material

We used ninety-six individuals, including the parental clones, of a diploid potato backcross population (CxE) (Celis-Gamboa et al., 2003). 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 (2002). All clones were grown in multi-year repeats in the field, Wageningen, The Netherlands during the normal potato-growing season in Tthe 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 for metabolomics, transcriptomic and proteomic analyses. The determination of carotenoids were as described in Kloosterman et al., 2010) and considered as a targeted metabolic analysis which includes compounds like zeaxanthin, violaxanthin and a compound for which the chemical behaviour is like violaxanthin (here denoted as ‘violaxanthin-like’)

Differential scanning calorimetry (DSC) is extensively used to study physical properties of starch granules. DSC-onset measurements (DSC onset) report water temperatures at which starch granules reach their gelatinization state (Kohyama and Sasaki 2006). DSC measurements provided gelatinization onset temperatures of starch granules for 96 individuals of the C x E population ranging between 61.5 and 66.7 ^{o}C. Enzymatic discoloration of tubers after peeling and exposed to air at room temperature in different time points such as after 5 minutes, 30 minutes, 3 hours and difference in discoloration between 3 h and 30 minutes and 3 hours is described in Werij et al., (2007). In this study, we considered only enzymatic discoloration after 5 min since the measurements at later time points are all highly correlated***.*** Potato tuber shape was scored between 1 (round) and 5 (long) (Celis-Gamboa 2002).

2.2 Microarray hybridizations and data processing

RNA was extracted from the 96 samples using the hot phenol method described previously (Bachem et al., 1996). All samples were labeled with both Cy3 and Cy5-dye using the low RNA input linear Amplification Kit, PLUS, Two colour (Agilent technologies) according to the manufacturer’s protocol starting with 2 μg of purified total RNA (see Kloosterman et al., 2012. for more detailed description). 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. We took into account only the Cy3 gene expression signals for further statistical analysis. For visualization, we used the gene nomenclature in the following way: Gene_Gene ID (for example: Gene_13945). The number refers to the gene ID of the supplementary material of Kloosterman et al., (2012)

2.3 LC-MS, GC-MS and proteomics data generation, processing and identification

Potato tuber samples were analyzed for variation in semi-polar metabolite composition using an untargeted accurate mass LC-MS approach. The untargeted metabolites are represented as centrotype_mass_scan number. For a hypothetical example: 818_795_918 means that the centrotype number is 818, the mass number 795 and the scan number 918. For visualization and simplicity we used LC-centrotype number (for example: LC_818). For more information on data generation, processing and identification see the materials and methods of chapter 3 of this thesis and/or Acharjee et al., (2011)

GC-MS and proteomics data (from 2D-DIGE experiment) were generated from the same 96 genotypes of the C x E population. Detailed materials and methods for GC-MS data generation, processing and identification are listed in Carreno-Quintero et al., 2012 or chapter 4 of this thesis. For visualization and simplicity we used GC-centrotype number (for example: GC_818). Detailed materials and methods for proteomics data generation, processing and identification are listed in the materials and methods section of chapter 5 of this thesis.

2.4 Random Forest regression

We used Random Forest (RF; Breiman 2001) for regression of the phenotypic traits: flesh colour, tuber shape, DSC onset and enzymatic discoloration after 5 minutes separately for the transcriptomics, metabolomics (LC-MS and GC-MS) and proteomics data sets, by using the randomForest package of R statistical software with default settings for the parameters. All four ~omics data sets were log2 transformed and then autoscaled (resulting in mean=0, sd=1 for each predictor variable). In each analysis, we estimated the variance explained by the RF model (R^2) on the out-of-bag (OOB) samples. This is essentially different from the R^2 for goodness-of-fit in normal ordinary least square (OLS) regression (Montgomery and Peck 1992). Variance explained (R^2) from RF is a value that is relevant for prediction of independent new samples (samples not used in the fitting of the statistical model), whereas the R^2 in ordinary least squares 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 (Breiman 2001). The greater the increase in the node purity values, the greater the importance of that particular variable (Breiman 2001). We used the variable importance values to rank the genes and metabolites with respect to their prediction for a quantitative trait. However, the standard RF procedure does not include tests for significance of the variables or of the model. Therefore we included a permutation test: the RF model was applied 1000 times for 1000 different permutations of the trait values and in each analysis we estimated the variance explained by the RF model (R^2) and variable importance of all variables in terms of decrease in node impurities. We ordered node impurity values from the permuted data sets and took the 95% percentile from the distribution of impurity values to assign a significance threshold for importance values of genes, proteins and metabolites. The same was done for the prediction R^2 values of the model: the 95%-percentile was used as a significance threshold for the prediction R^2 value in RF regression (Acharjee et al., 2011).

2.5 QTL analysis and cis- and trans-eQTLs

We mapped expression QTLs (eQTLs) from the gene expression data, metabolite QTLs (mQTLs) both for LC-MS and GC-MS, and protein QTLs (pQTL) from the proteomics data to find regions on the genome explaining genetic variation in gene expression, metabolites and proteins values using the integrated linkage map of the C and E parents for QTL analysis (Kloosterman et al., 2012).

Further, we used the potato genome physical map (Xu et al., 2011) to investigate eQTL physical positions of genes identified as predictive for phenotypic traits from the Random Forest analyses. The potato oligo (60-mer) microarray (POCI) used in the experiments contains 42,034 features based on a potato unigene set (Kloosterman et al., 2008). To allow discrimination between cis- and trans-eQTLs all unigenes were blasted against the genome scaffold sequences, predicted coding sequences (CDS) and predicted gene regions (including 5' and 3'-UTR's). Features with a unique and significant hit were assigned to genome scaffolds for which the majority has chromosome information. Identified QTLs on the same linkage group as their physical map position are identified as cis-acting while QTLs on different linkage groups are defined as trans-acting. Features on the array for which no physical map position could be assigned are classified as unknown (Kloosterman et al., 2012).

2.6 Network reconstruction

From the RF analyses, for each of the traits, we obtain a list of significant genes, metabolites and proteins. For visualization in a correlation network, we reduced the number of genes by checking chromosomal positions of each of the predicted significant genes, metabolites (LC-MS and GC-MS) and proteins. We took the most significant gene, LC-MS, GC-MS or /and proteins per chromosome as representative for that data set and chromosome and made a Pearson correlation network with these genes, metabolites, proteins and traits as nodes in the network (Yuan et al., 2008) and the correlation coefficient as their strength of the interaction (as edges in the network). A significance threshold for the correlation coefficient (α=0.01) was used to draw lines between genes, metabolites, proteins or traits.

3 Results

3.1 Selection of ~omics features predictive for quality traits

The prediction of variation in potato tuber flesh colour from the gene expression, LC-MS, and protein data sets was quite high (> 50% explained variance using all features, 60% to 75% for smaller subsets of only significant features), but much lower for the GC-MS data (10% and 33% for unselected and selected features (Table 1 and 2). For flesh colour, the microarray data and the LC-MS data were equally good for prediction in terms of the explained variance, but the numbers of significant features were very different: the prediction of flesh colour using 7 significant LC-MS features is almost the same as for 233 significant gene expression profiles (of which the genes are distributed over different regions in chromosome) from the microarray data set. From the gene expression data, the gene which ranks first and third with respect to variable importance for predicting flesh colour was a beta-carotene hydroxylase (Bch). Two oligos were present on the array targeting the same *beta-carotene hydroxylase* gene, hence the two high ranks for the same gene. Another gene from the carotenoid pathway, *zeaxanthin epoxidase (Zep)* ranked forty-fourth. Based on our current knowledge of potato tuber flesh colour and carotenoid content (Kloosterman et al., 2010; Wolters et al., 2010), these two genes were expected to be associated with flesh colour. From the GC-MS data, out of six significant metabolites, four had an annotation: malic acid, 2-4-5-trihydroxypentanoic acid, glucopyranose, 2-butenedioic acid(z)- and bis(trimethylsilyl) ester.

For the three other traits the gene expression data was more predictive than the metabolomics and proteomics data sets, and for tuber shape actually no significant features were found for GC-MS, LC-MS and the protein data, while significant gene expression differences explain 55% of the variation in tuber shape. For starch gelatinization (DSC) no LC-MS features were significant, for enzymatic discoloration no GC-MS features were significant.

Table 1: Percentage explained variance (R^2) in out-of-bag (OOB) prediction by Random Forest (RF) models using all genes, LC-peaks, GC-peaks or proteins, separately. Non-significant models indicated as NS (permutation p-value > 0.001).

Quality traits	**Gene expression**	**LC peaks**	**GC peaks**	**Proteins**
Flesh colour	58%	63%	10%	53%
Tuber shape	32%	NS	NS	NS
DSC Onset	42%	NS	12%	22%
Enzymatic discoloration	14%	16%	NS	13%

3.2 Associated genomic regions

Using Random Forest regression, we generated lists of candidate genes, metabolite features and proteins that are predictive for quality traits of potato tubers (Table 1,2). However, from this prediction we cannot conclude that these associations necessarily have a basis in genetic differences between the plants since they could also be caused by environmental variation in both the trait and the levels of the features, or in developmental differences. Therefore we also investigated QTL positions for both the phenotypic traits and the ~omics features. When cosegregation of an eQTL, mQTL or pQTL with a trait QTL is observed, this could imply a functional relationship or identify causal genes/proteins/metabolites; however, this is not necessarily the case since any linked but functionally unrelated QTLs will also show this cosegregation. Still, these sets of predictive genes highlight genomic regions of interest, comparable to standard QTL analysis, and can provide additional information or narrow down the region of the causative polymorphism

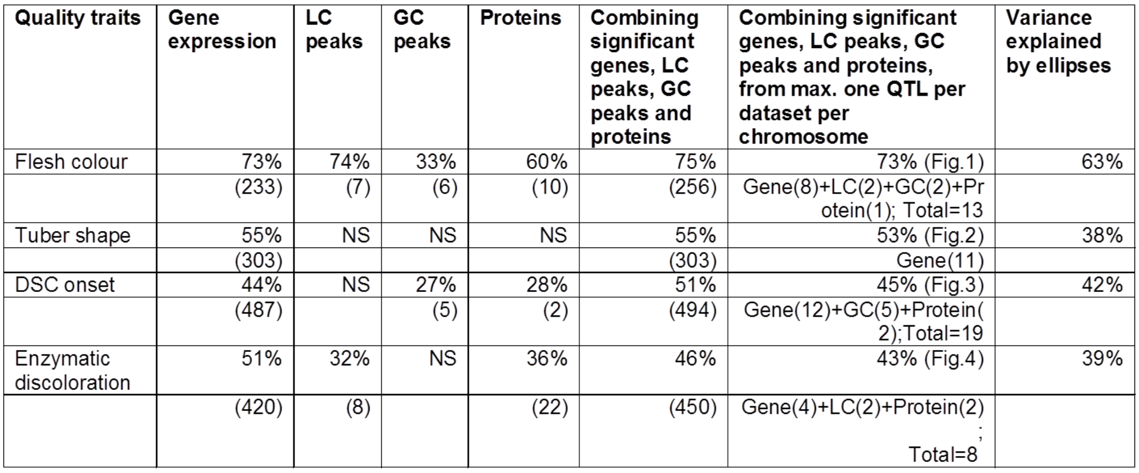
For flesh colour the 233 eQTLs of significantly predictive gene expression profiles from RF mapped to eight chromosomes: 2,3,4,5,8,9,11 and 12; large numbers (132) of eQTLs were mapped to chromosomes 2 and 3. Significant GC-MS peaks mapped to two chromosomes, 1 and 2; QTLs for significant LC-MS peaks mapped to chromosomes 2 and 3 and QTLs for significant protein spots mapped to chromosome 3. Using 13 variables (one per chromosome with a QTL, per data set) as predictor set for flesh colour the RF prediction explains 73 % of the phenotypic variation. In Fig. 1, chromosomes that have QTLs across multiple ~omics data sets are indicated with ellipses. For flesh colour, the 3 chromosomal regions with QTLs across multiple data sets are predicted to explain 63% of the phenotypic variation.

For tuber shape the expression QTLs of the significant genes mapped to chromosome 1 to 11 but not on chromosome 12. However, the largest number (185) expression QTLs of the significant genes map to chromosome 10. Using one representative gene from each of those eleven chromosomes as a predictor set for tuber shape, the RF prediction explains 53 % of the variation in the trait. No proteins or GC- or LC-variables were included since none were significantly associated to tuber shape. Only the chromosome 10 representative gene explains already 38 % of the phenotypic variation.

For starch gelatinization (DSC onset), the eQTLs of the significant genes mapped to all the twelve chromosomes but the largest number (201) mapped to chromosome 2. QTLs for significant GC-MS peaks mapped to 5 chromosomes: 2, 4, 5, 9 and 11. QTLs for significant proteins were mapped to chromosomes 2 and 5. These 19 selected variables explain 45 % of the variation in starch gelatinization. In Fig. 3, chromosomes that have QTLs across data sets are indicated with ellipses; for starch gelatinization (DSC onset) the 4 ellipses with QTLs across data sets (corresponding to 4 chromosomal regions) are predicted to explain 42% of the phenotypic variation for DSC.

For enzymatic discoloration the expression QTLs of the significant genes mapped to four chromosomes: 1,3,5 and 8; QTLs for significant LC-MS peaks mapped to chromosomes 3 and 5 and for significant protein spots to chromosomes 1 and 3. These 8 selected variables (genes and LC-MS peaks) explain 43 % of the variation in enzymatic discoloration. In Fig. 4, chromosomes that have QTLs across data sets are indicated with ellipses; for enzymatic discoloration 3 ellipses are predicted to explain 39% of the phenotypic variation.

Table 2: Variance explained (R^2) using only significant gene expression, LC-peaks, GC-peaks or proteins. The numbers of significant (permutation p-value < 0.001) genes, LC-peaks, GC-peaks or proteins are between brackets. NS: no significant relationship was found



Quality traits	Gene expression	LC peaks	GC peaks	Proteins	Combining significant genes, LC peaks, GC peaks and proteins	Combining significant genes, LC peaks, GC peaks and proteins, from max. one QTL per dataset per chromosome	Variance explained by ellipses
Flesh colour	73%	74%	33%	60%	75%	73% (Fig.1)	63%
	(233)	(7)	(6)	(10)	(256)	Gene(8)+LC(2)+GC(2)+Protein(1); Total=13	
Tuber shape	55%	NS	NS	NS	55%	53% (Fig.2)	38%
	(303)				(303)	Gene(11)	
DSC onset	44%	NS	27%	28%	51%	45% (Fig.3)	42%
	(487)		(5)	(2)	(494)	Gene(12)+GC(5)+Protein(2);Total=19	
Enzymatic discoloration	51%	32%	NS	36%	46%	43% (Fig.4)	39%
	(420)	(8)		(22)	(450)	Gene(4)+LC(2)+Protein(2); Total=8	

3.3 Integration of ~omics data and network visualization

With Random Forest we selected genes, metabolites and proteins that have a significant association with quality traits. Many of these have QTLs on the same genomic regions and can be considered as redundant. Therefore we selected a single representative feature per QTL region per data set for network visualization (Fig. 1-4). The networks show Pearson correlation coefficients when these were significant at a significance level of 0.01. The nodes in the network show the phenotypic trait of interest and selected genes, metabolites and proteins (one per QTL per data set). Positive correlations are shown in solid lines, negative correlations in dotted lines. Chromosomes that have QTLs across multiple ~omics data sets are indicated with ellipses (considered as clusters) with bigger node size of the genes, proteins or metabolites

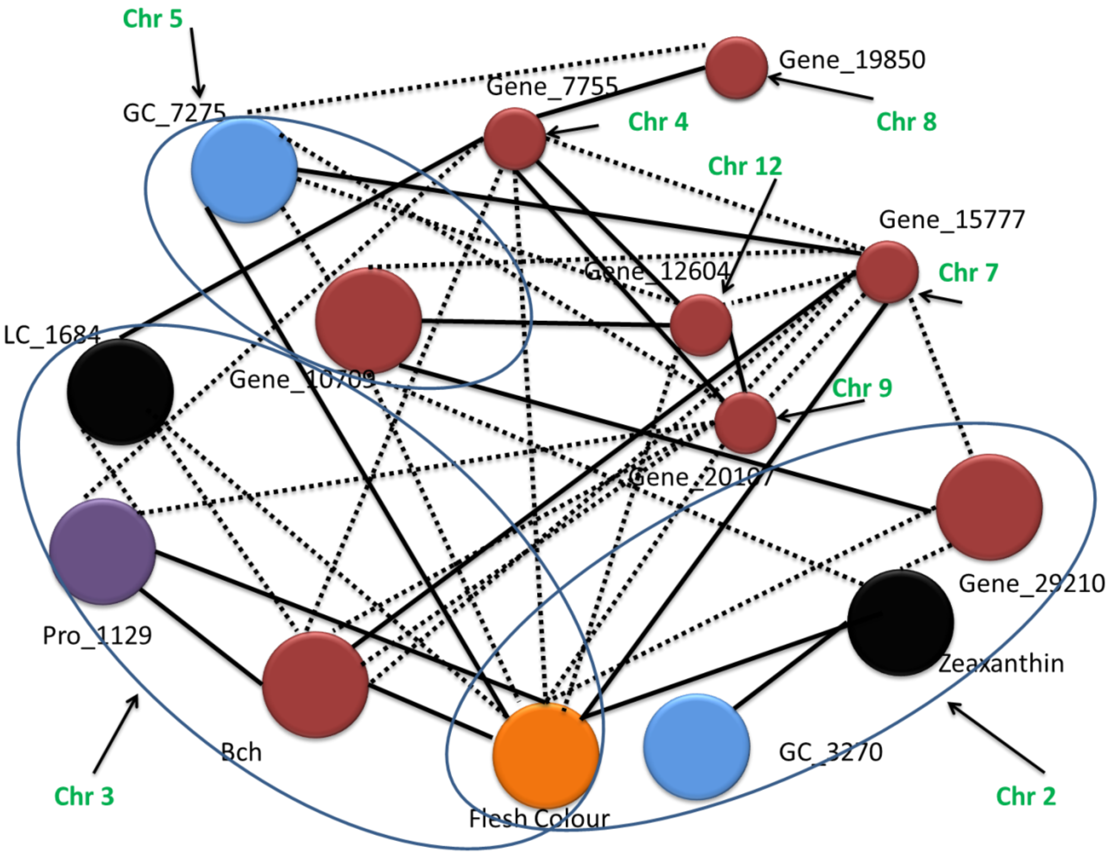


Figure 1: A Pearson correlation network of gene expression features (red), metabolites from LC-MS (black), metabolites from GC-MS (light blue), proteins (dark blue) and the phenotypic trait tuber flesh colour (yellow). The dotted lines represent negative correlation coefficients, solid lines represent positive correlation coefficients. Only correlation coefficients significant at α=0.01are considered. Chromosomes with QTLs across multiple ~omics data sets are shown in blue colour in elliptical form. Bch = beta-carotene hydroxylase, LC_X =represents metabolites derived from LC-MS with centrotype number(X), GC_X = represents metabolites derived from GC-MS with centrotype (X), Gene_X= Gene with gene ID (X)

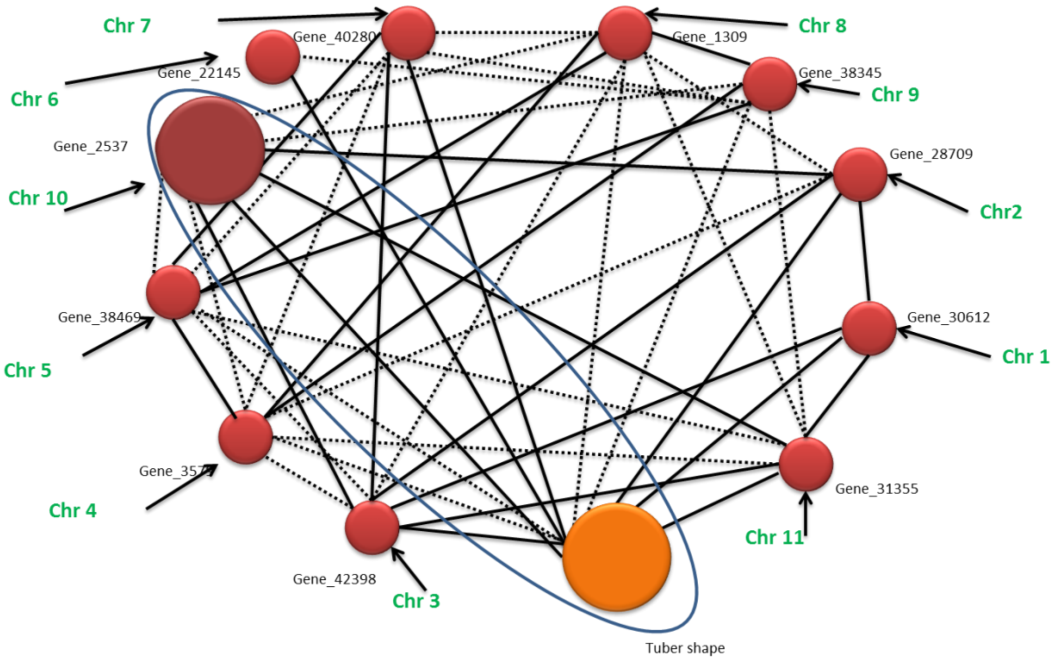


Figure 2: A Pearson correlation network of genes (red) and tuber shape (yellow). The dotted lines represent negative correlation coefficients; solid lines represent positive correlation coefficients. Only correlation coefficients significant at α=0.01are considered. Gene_X= Gene with gene ID (X), Chr. No X = Chromosome number

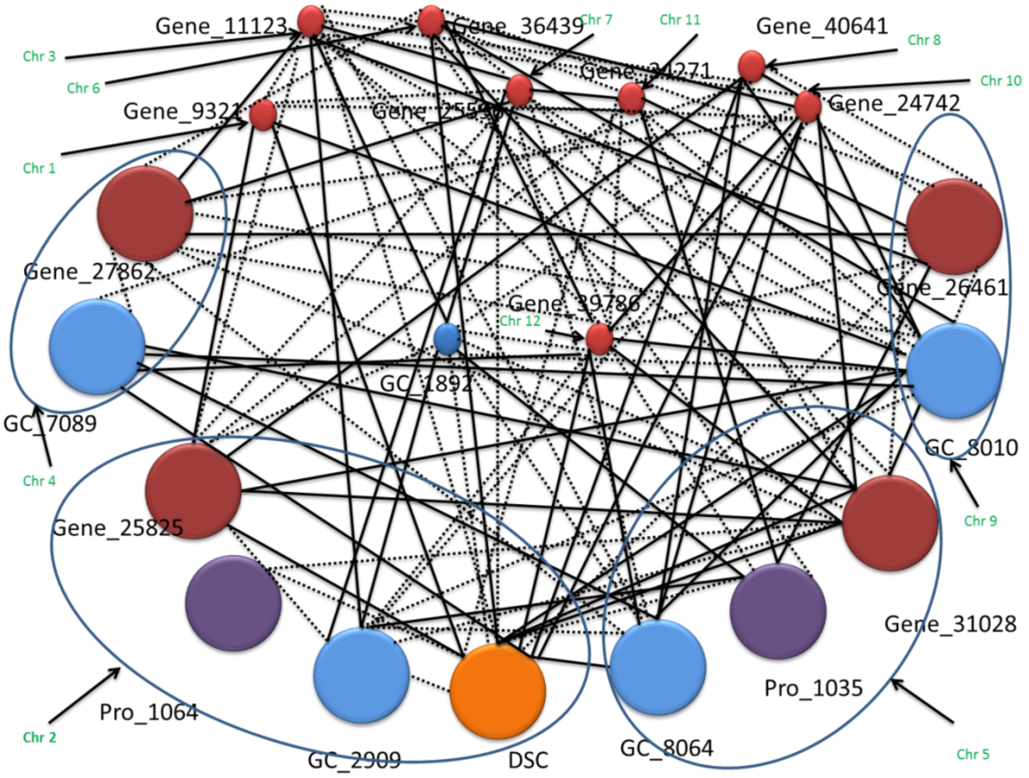


Figure 3: A Pearson correlation network of genes (red), metabolites from GC-MS (light blue), proteins (dark blue) and DSC onset (yellow). The dotted lines represent negative correlation coefficients, solid lines represent positive correlation coefficients. Only

correlation coefficients significant at α=0.01are considered. Clusters are shown in blue colour in elliptical form. GC_X = represents metabolites derived from GC-MS with centrotype (X), Gene_X= Gene with gene ID (X), Chr. X = Chromosome number

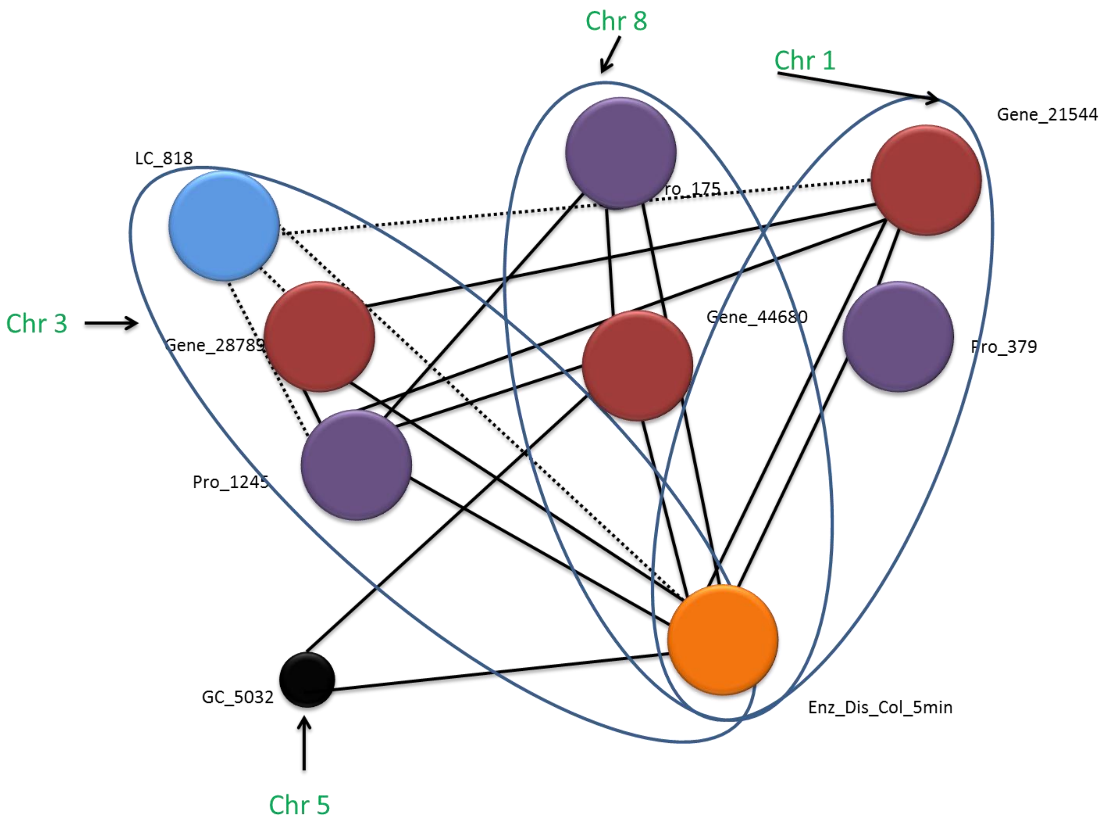


Figure 4: A Pearson correlation network of genes (red), metabolites from LC-MS (black), proteins (dark blue) and enzymatic discoloration(yellow). The dotted lines represent negative correlation coefficients, solid lines represent positive correlation coefficients. Only correlation coefficients significant at α=0.01 are considered. Clusters are shown in blue colour in elliptical form. LC_X =represents metabolites derived from LC-MS with centrotype number(X), Gene_X= Gene with gene ID (X), Chr. X = Chromosome number

4 Discussion

We used Random Forest regression for integrating transcriptomics, metabolomics and proteomics data for prediction of four quality traits of potato: tuber flesh colour, DSC onset, tuber shape and enzymatic discoloration. For each of these traits, we selected sets of genes, metabolites and proteins that were significant in explaining variation in the trait. We quantified the amount of variance explained in prediction using these selected sets of ~omics features, and we constructed correlation networks for subsets of genes, metabolites and proteins using QTL mapping information.

4.1 Flesh colour

For tuber flesh colour beta-carotene hydroxylase (Brown et al., 2006) and zeaxanthin epoxidase (Zep) (Wolters et al., 2010) were ranked first and forty-fourth respectively both of which have previously been associated with flesh colour in potato tubers (Brown et al., 2006). A more detailed discussion of these findings is presented in the discussion section of chapter 3.

In GC-MS, six metabolites were significant and out of these six, four were annotated. It was observed that one of the compounds was identified as malic acid. Although there is no direct link between tuber flesh colour and malic acid, Ruiz and Egea (2008) reported correlation of malic acid to skin colour in apricot. For the other three metabolites we did not find any connection with carotenoids or with flesh colour.

Combining all the significant genes, LC-peaks, GC-peaks and proteins, the OOB variation explained (R^2) was 75%, only slightly more than what gene expression or LC-MS data explain by themselves (Table 2) which indicates that there are correlations among the variables across data sets. For example: from proteomics data, Pro_250 and from LC-MS data, 1710_640_1762 is highly correlated with beta-carotene hydroxylase gene expression with Pearson correlation coefficients of 0.88 and 0.72 respectively.

4.2 Tuber shape

For tuber shape regressed on the gene expression, LC-MS, GC-MS and proteomics data sets separately, only expression data was found to explain significant variation. More than half of the eQTLs are mapped on chromosome 10. Tuber shape is thought to be regulated by a single locus Ro on chromosome 10, where round (Ro_) is dominant over long (roro) (Van Eck et al., 1993b, 1994; Jacobs et al., 1995). At the Ro-locus a series of multiple alleles can explain all intermediate shapes between round (going to flat) and long (Van Eck et al., 1994). The large number of eQTLs of genes predictive for tuber shape can be due to linkage to this Ro locus. No metabolites and proteins were found as significantly predictive of tuber shape, which suggests that for this particular trait, gene expression data is more informative than metabolomics or proteomics data.

4.3 DSC onset

For DSC onset, we found significant gene expression, metabolite levels (GC) and proteins that are associated with the trait. Using those 19 significant variables, the variation explained was 45%. In order to find what genomic regions the selected genes are regulating, eQTL analyses were performed. The analyses showed many associations with genomic regions in chromosome 2 with also the highest explained variation compared to other chromosomes. For gene contig MICRO.9632.C4 the eQTL analysis revealed a large

QTL region on the last portion of chromosome 2 between 73 cM and 86 cM. eQTL analysis for EST BF_LBCHXXXX_0013B11_T3M.SCF produced a single QTL on the same chromosomal location as the QTL found with the QTL analysis similar eQTL was also found for gene contig MICRO.13279.C1. For GC-MS out of five metabolites we could identify three: proline, glucopyranose and 2-Piperidone. The proteins which were associated with DSC onset could not be identified.

4.4 Enzymatic discoloration

Transcriptomics and metabolomics analysis on enzymatic discoloration after 5 minutes resulted in 420 significant genes and 8 significant metabolites, among which two were putitatively identified as caffeoylquinic acid methyl ester and tyrosine (Werij et al. 2007). A more detailed discussion of these findings is presented in the discussion section of chapter 3. We used RF regression for integrating the transcriptomics, metabolomics and proteomics data sets with phenotypes of interest. This procedure can handle high-dimensional data and this method has an internal cross-validation (using the OOB samples, see Breiman 2001). However, other variable selection methods like for example: LASSO, Elastic net or ranking methods for example: ridge regression, Partial least squares regression also can be applied. For more details on the different statistical methods and comparison of their performance, see chapter 2. One of the limitations of this approach is that a RF model by default will (or, at least, can) use all variables simultaneously and no automatic variable selection is included. If we want to perform variable selection, we need to select variables based on a significance criterion or include a variable selection procedure similar to the one that Diaz-Uriarte et al. (2006) used for classification with Random Forest.

In this paper we used genetic information through QTL analysis on the one hand and prediction of the traits using RF analysis from transcriptomics, metabolomics and proteomics analysis on the other hand. From the QTL analyses, we can identify the map position of the QTLs for gene expression or metabolite signals but we expect that functional genes and genes that are only linked and that influence other pathways will show similar correlations (For example: we get lots of genes for flesh colour in chromosome 3 but we know that only Bch is responsible for flesh colour in chromosome 3, so probably the remaining genes are associated because of linkage).

In RF regression prediction of phenotype from metabolomics, transcriptomics or proteomics data is possible in a way that genes, metabolites and/or proteins might be linked with a phenotype but independent of the genetic information (Acharjee et al., 2011). In this study, the predictive genes are co-localized with trait QTLs. Further, for the predictive genes we checked where they are located in the physical position of the genome. That also means that, given a genome sequence and predictive gene list associated to the trait of interest, it can be useful to find genomic positions of those genes irrespective of QTL analysis.

The key advantage of eQTL, mQTL or pQTL mapping in addition to the traditional mapping of phenotypic QTLs, is that it connects variation at the level of RNA expression, metabolite or protein abundance to variation at the level of DNA. The latter provides versatile tools for breeding whereas the first can reveal information on the biology of a trait and can direct to new candidate genes. Mapping of eQTL to the gene itself indicates that cis changes are responsible for the different expression levels, whereas mapping positions of eQTL different from the position of the corresponding genes indicate trans-regulation which allows deriving regulatory networks of genes (Kloosterman et al., 2010).

In this chapter RF regression was used as a tool for data integration of metabolites, gene expression and protein profiles relating to a phenotypic trait of interest where it was used to identify leads for further exploration. For example: flesh colour was associated with metabolites and after putative identification of the metabolic peaks 4,7-Megastigmadiene-3,9-diol-glucoside and 2,3-Dihydroxy-4-megastigmen-9-one-glucoside were identified as carotenoid derived compounds. Such approaches give us leads for further research on the metabolites and help in hypothesize which components (genes, metabolites, proteins) are in a specific pathway of interest and the genetic basis of the genes, metabolites or proteins involved in the pathway. Metabolite peaks that are not identified but that also show an association to the trait, could in many cases be breakdown products carotenoid pathway (Kloosterman et al., 2010).

In this study, we made a strategy to integrate multiple ~omics data with traits of interest and select the number of co-expressed genes, metabolites and proteins. Prediction of significant genes was obtained through RF regression, then through a genetical genomics study (Jansen and Nap (2001) we mapped those QTLs from genes (eQTLs), metabolites (mQTLs) and proteins (pQTLs). We selected one gene, metabolite and/or proteins per chromosome if they mapped to the same position to use further for integrated network analysis and this can then be subsequently used as predictive network for the trait of interest. By doing so, we selected genes which are not only predictive for a trait of interest but also explain variation of the trait of interest. For example: in flesh colour, a beta-carotene hydroxylase gene ranks first in the predictive gene list and also explains around 50 % of the trait variation in chromosome 3. This indicates that it is possible to select possible candidate genes or a candidate genetic or metabolic pathway involved in a trait of interest or at least give a lead for further study of these associated features. Combining both QTL (eQTL, mQTL and pQTL) analysis and prediction of traits using RF, or other regression techniques suitable for ~omics data (Acharjee et al., 2011), gives us a clue about the candidate genes, metabolites or proteins which are associated with the phenotype but also the genetic information about those genes, metabolites and proteins from QTL analysis.

Within individual ~omics data sets: LC-MS data for flesh colour trait results in 63 % variation (R^2) explained of the flesh colour which is the highest variation explained compared to the other traits. So, for flesh colour LC-MS data is more informative than other ~omics data sets. Gene expression and proteomics data for flesh colour is explaining 58 % and 53 % of the variation.

For other traits such as DSC onset, tuber shape and enzymatic discoloration gene expression data seems to be more informative as a higher amount of variation is explained than for metabolomics and proteomics data. Even in the case of DSC, using only the expression data gives already the highest percentage of explained variance compare to the other traits.

For selected combined data sets the results are interesting because there is an improvement of the R^2 prediction value as compared to the unselected data. This improvement is most likely due to filtering out the noise variables from the data set which is explained in Acharjee et al. (2011) in more detail. For DSC onset, none of the LC-MS peaks were significant which might indicate that primary metabolism is more important for this trait compared to secondary metabolism, whereas for enzymatic discoloration the situation is the other way around.

Now questions arise which type of ~omics data is important to study a trait? From our study it is clear that generation of ~omics data largely depends on the trait and biological information on the trait of interest. For example, if the trait is a quality trait then metabolomics data will be useful to investigate further the biochemical pathways and arise at potential candidate genes (for example: flesh colour was associated with metabolites and after putative identification of the metabolic peaks 4,7-Megastigmadiene-3,9-diol-glucoside and 2,3-Dihydroxy-4-megastigmen-9-one-glucoside were identified as carotenoid derived compounds). But given any trait, gene expression data seems to be very useful anyway, to check the variation of the gene expression level linked with the trait. However, if we do not have any prior information about the trait of interest (novel trait), it will be difficult to say what type of ~omics experiments will be useful and give the highest chance to arrive at new candidate genes.

We selected significant genes, metabolites and proteins based on permutation tests and finally selected representative individual peaks for networks based on eQTL, mQTL or pQTL information. Network analysis was done to interpret how a particular trait is associated with gene expression, metabolite and protein data. Also in the Fig.1-4, in the ellipses, we indicated genes, metabolites, proteins and traits by large size which contributes to the large variation of the phenotypic trait. Those genes, metabolites and proteins might consider as lead with the connections to the phenotype. Although, the expanded potential chromosomal regions do not lead automatically to genes or metabolites directly involved with the trait, however, it might be helpful to know more about the metabolic pathway and indirect acting genes and/or metabolites. Further study would be needed to analyze the combined effect of multiple QTL regions over different chromosomes.

All in all it can be concluded that although this approach has some large limitations (mostly with regards to absence of prior knowledge regarding genes, metabolites or proteins of the trait under investigation, and difficulties in identification) the fact that networks can be made holds some promise for the future. What needs to be further worked out are conclusive ways to validate predictive networks. For practical breeding purposes it has become clear that the network approach is still too complex to deal with at this moment for most crops however for some traits (like enzymatic discoloration) with a limited number of chromosomes For now it seems that the biggest step forward with using these approaches could be to zoom in on the biology of traits and identify new candidate genes for traits.

Chapter 7

General Discussion

1 Introduction

In recent years the rise in high-throughput technologies has created a tremendous impact on different domains of science, among which also plant biotechnology, plant breeding and genetics. These high-throughput technologies, such as microarrays, RNA-Seq and mass spectrometry, generate very wide data sets consisting of gene expression, metabolite intensity or protein expression data of very large numbers of genes, metabolites or proteins but usually measured on relatively few (50-200) samples. These 'omics' technologies are giving researchers new tools to help the identification of the genetic underpinnings of crop improvement, for example the genes that contribute to improved productivity and quality of modern crop varieties. These ~omics technologies enable identifying the factors affecting crop growth and yield, and provide the data that can be utilized to investigate the complex interplay between the plant, its metabolism, and also abiotic or biotic stresses.

In this thesis, we investigated different ~omics data sets: transcriptomics, metabolomics (LC-MS and GC-MS) and proteomics in a segregating population (C x E) of a diploid potato cross (Celis-Gamboa 2002). In addition we had the availability of molecular marker and phenotypic data sets from the same potato mapping population (Celis-Gamboa 2002) as well as a reasonably dense genetic map (Anithakumari et al., 2010) and sequence information (Kloosterman et al., 2012).

The main objectives of this thesis were the following: 1) to evaluate methodology that can help predict phenotypic traits from wide ~omics data sets, allowing for the number of ~omics variables (genes, metabolites or proteins) being much larger than the number of the samples, and allowing for correlations between the variables and taking into account that maybe only a small number of variables will be involved in the trait. 2) Achieve integration across multiple platforms such as transcriptomics, metabolomics and proteomics so that combined they can predict a particular phenotype of interest, with the aid of genetics (genetical genomics, eQTL, pQTL, mQTL analyses).

We divided our objectives into the following parts and results were presented in different chapters: chapter 1 introduces the scientific interest of relating phenotypic traits to ~omics data sets and describes the problems that need to be overcome and the tools that are needed. In chapter 2 we compared statistical methods suitable for prediction of a quantitative phenotypic trait from an ~omics data set, considering aspects such as correlations among the variables, and ranking of genes, metabolites or proteins for the prediction of the trait. Chapter 3 was about to find a strategy to integrate transcriptomics and metabolomics (LC-MS) data sets and select a subset of the metabolites and transcripts which show an association with phenotypic traits like for example flesh colour and enzymatic discoloration. In chapter 4, we used gas chromatography (time-of-flight) mass spectrometry (GC-TOF-MS) data sets to identify genetic factors underlying variation in primary metabolism in the C x E mapping population. We performed a QTL analysis for starch and cold sweetening related traits and inferred links between these phenotypic traits and primary metabolites. In chapter 5, we performed a proteomics analysis of potato tubers in order to obtain an insight into the relationships between protein and quality traits such as enzymatic discoloration, starch and cold sweetening related traits.

In chapter 6, an integrated analysis with multiple ~omics data sets such as transcriptomics, metabolomics and proteomics was done. This chapter (chapter 7) provides a general discussion on all the findings. Suggestions for future approaches for the dissection of traits and ~omics data sets are given.

One of the major goals in potato breeding is high yield. In addition to yield, an improved agronomic performance and quality is desirable. Depending on the target market, the focus on quantity is shifting more towards breeding for quality. For instance, from a practical grower's perspective, being able to produce outstanding quality (crisps, French fries, industrial use of starch etc.) can offer a distinctiveness that can create extra revenue (Caswell and Mojduszka 1996). For the consumer market, some physical properties connected to quality traits, like tuber shape, tuber colour and size, are easily observable by consumers and hence major factors determining market value of a product.

In this study, we considered quality traits in potato such as tuber flesh colour, enzymatic discoloration, phosphate content, cold sweetening, tuber shape and starch gelatinization. These are traits for which in many cases there was no prior knowledge with respect to which genes might regulate or determine these traits. Identifying metabolites or proteins from ~omics data might help in getting an idea about the potential genes involved.

In this thesis, we made an attempt to integrate different types of ~ omics data with phenotypic traits of interest.

We considered different approaches:

- Relating a phenotypic trait to a single ~omics data set (whether transcriptomics, proteomics, LC-MS or GC-MS)
- Relating a phenotypic trait to multiple ~omics data sets simultaneously

2 Relating a phenotype to an ~omics data set

~Omics data generated from microarray, LC-MS and GC-MS are commonly used to study the behaviour of genes, proteins and metabolites, and typically generate large data sets. Here 'large' refers to the number of genes, proteins, metabolite peaks compared to the much smaller number of samples (individuals). This situation with a much larger number of variables (p) than samples (n) is often referred to as the p>>n problem. There will be collinearity due to p>>n (Kiers and Smilde 2007) but also because of high correlations due to common biological functions. In chapters 2 and 6, we used a microarray with gene expression data of around 15,000 genes measured on around 90 progeny individuals (from a statistical point of view, these are the experimental units, n=90). In plant breeding research, the phenotypic data could be disease scores, growth characteristics, agronomical traits or quality traits such as flesh colour of potato, phosphate content etc. Such phenotypic data can be scored on a continuous scale, an ordinal scale or as binary scores. To link such phenotypes with ~omics data one could think of a regression approach (In case of binary response, one could use a classification or a logistic regression approach) where the phenotypic trait is considered as a response variable and an ~omics data set as a predictor set. This makes sense as we are usually interested in the phenotypic trait as predicted from the molecular profile variation between individuals in a population. Due to p>>n and collinearity among the set of variables we cannot invert the variance-covariance matrix to estimate regression coefficients and hence traditional statistical methods such as multiple linear regression can not be applied. However, univariate regression analyses per variable are possible; but even then, due to the large number of variables, we have to take precautions to avoid a large number of false positive test results. One of the ways to deal with such false positives is to apply an FDR correction to the results of the univariate analyses (Benjamini and Hochberg 1995). This univariate approach is possible even though genes, proteins and metabolites are expected to act in a combined manner rather than each by itself in a biological system. For an approach in which we consider more ~omics variables together, we need methods that can be used despite the overabundance of candidate variables associated with a response variable. One of the ways is to apply regularized multiple regression methods or machine learning methods. In chapter 2, we applied and compared eight different regression methods to predict potato flesh colour from a metabolomics (LC-MS) data set in a potato mapping population. The objective was to evaluate methods which can handle p>>n and high correlations among predictor variables (here the metabolic peaks). We compared these methods in terms of mean square error of prediction (MSEP), goodness of fit (R2), variable selection properties and the ranking of the variables.

In this thesis, we did not have an independent data set for validation and hence we applied cross validation. Cross validation can be used for two purposes 1) optimization of hyper parameters (for example : number of components in the PLS model) in a predictive model by using the prediction error as a criterion to choose values for the hyper parameter(s) of these models. 2) Cross validation is also used for estimation of the prediction error (usually called as mean square error of prediction (MSEP)) of a method, so, to validate and quantify the predictive quality of a procedure.

One of the ways to do cross validation is partitioning a data set into a number of complementary subsets (this number is denoted as k), performing the analysis on one subset (called the training set or learning set), and validating the analysis on the other subset (k-1, called the validation set or test set). For example, the total number of samples can be divided into k=3, k=5, or k=10 subsets. In this thesis we used 10-fold cross validation (k=10). The schematic diagram of cross validation is shown in Figure 1. To reduce variability, multiple rounds of cross-validation are performed using different partitions, and the validation results are averaged over these rounds.

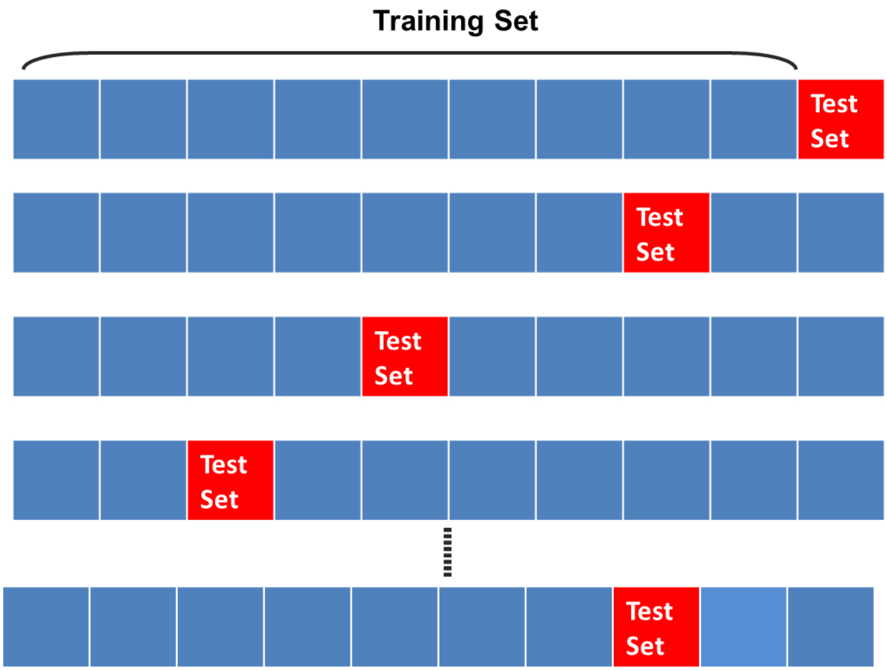


Figure 1: A 10-fold cross validation scheme is shown considering a “wide data” set where samples are in rows and variables (genes, metabolites or proteins) are in columns. The blue colour partitions grids are acting as training/learning samples whereas the red ones are acting as test data or “Test Set”. The figure shows one possible partition into 10 folds but many different possible partitions can be used.

The mean square error of prediction (MSEP) is usually used for the quality of prediction performance on an independent test set and estimate the error of prediction. Often, a (large enough) test set is not available. In such situations, the MSEP has to be estimated from the learning data, i.e., the data used to train the regression. The mean squared error of prediction (MSEP) is obtained by averaging the squared prediction errors (differences between observed and estimated values) of the test samples. A lower MSEP value corresponds to a better predictive model.

When comparing MSEP values for the different methods to predict flesh colour from the LC-MS data, we found that elastic net (EN) had the lowest MSEP corresponding to a better predictive model than ridge regression (RR), least absolute shrinkage and selection operator (LASSO), support vector machine (SVM), principal components regression (PCR), partial least squares (PLS), random forest (RF) and sparse partial least squares regression(SPLS). RR, PCR and PLS had similar MSEP values for our data set. The performance of the methods could depend on the type of data, size of the data set, signal to noise ratio in the data set under study and hence it is difficult to generalize which is the “best” method. For example, Xu et al., 2012 also explored several regularized regression methods: ridge regression (RR), LASSO, principal components regression (PCR), partial least squares (PLS) regression and sparse PLS (SPLS) to effectively identify associations between a genetic region and a continuous trait. In their situation they found LASSO to perform better compared to the other methods in terms of the power.

Contrary to what would have been expected from literature (Tibshirani 1996) we found that the method LASSO selected a group of correlated variables, whereas, LASSO should, according to Tibshirani 1996, have selected only one variable from a set of correlated variables So, if we really want to select only a single variable from a set of correlated variables, LASSO will not be ideal to apply. In metabolomics data we often find metabolic peaks that are highly correlated (r>0.9) for example as a result of fragmentation of a compound during spectrometry. In such situations, to reduce redundant peaks out of many correlated ones requires a method which will allow to select one variable from the group of correlated variables (metabolic peaks). If, on the other hand, we are interested in selecting groups of correlated variables, we found that SPLS in our data set performed better than EN in selecting a larger number of correlated variables.

In our analyses in chapter 2, we observed that variable selection methods rather than non-selection methods performed better in terms of the MSEP. This could be due to the fact that the variables which are not associated with the phenotypic trait (noise variables) get regression coefficients with the value zero, so that they effectively drop out of the regression model. Ogutu et al., 2012 evaluated the predictive performance of six regularized linear regression methods– ridge regression, ridge regression BLUP (best linear unbiased prediction), LASSO, adaptive LASSO, elastic net and adaptive elastic net– for predicting marker effects using dense SNP markers in the context of genomic selection (Meuwissen et al., 2001; Heffner et al., 2009; Jannink et al., 2010; Heslot et al., 2012). They found that all the six models had relatively high prediction accuracies for their simulated data set, but there also accuracy was higher for the LASSO type methods than for ridge regression and ridge regression BLUP.

3 Relating a phenotypic trait to multiple ~omics data sets

If we have obtained multiple ~omics data sets and want to relate them to a trait of interest, we are in a different situation (multiple ~omics data) than just relating each of them separately to the trait (as explained in the previous section). The aim is to find relationships among these multiple different levels of regulation with a phenotype of interest. In this case, for example, potato tuber flesh colour, vs. gene expression and metabolite levels. Given multiple ~omics data sets and such a specific phenotypic trait, one could fuse two (or more) ~omics data sets into a single data set and use them as one large ~omics data set so that we can apply the above mentioned methods again. However, due to different error structures within the data sets, this might not always give the best results. For example, it could happen that due to these different errors, or to possibly very different scales of the data sets, variables from one data set are preferred over variables from the other data set(s), while we are actually maybe interested in obtaining predictive variables across the data sets.

In chapter 3, we developed a strategy to relate a quality trait (tuber flesh colour of potato) to LC-MS and transcriptomics data sets using RF regression. The strategy we followed was: treat the ~omics data sets separately and regress a phenotype separately on each one, based on these regression results select subsets of variables for each and then evaluate the associations among the variables selected from the different data sets.

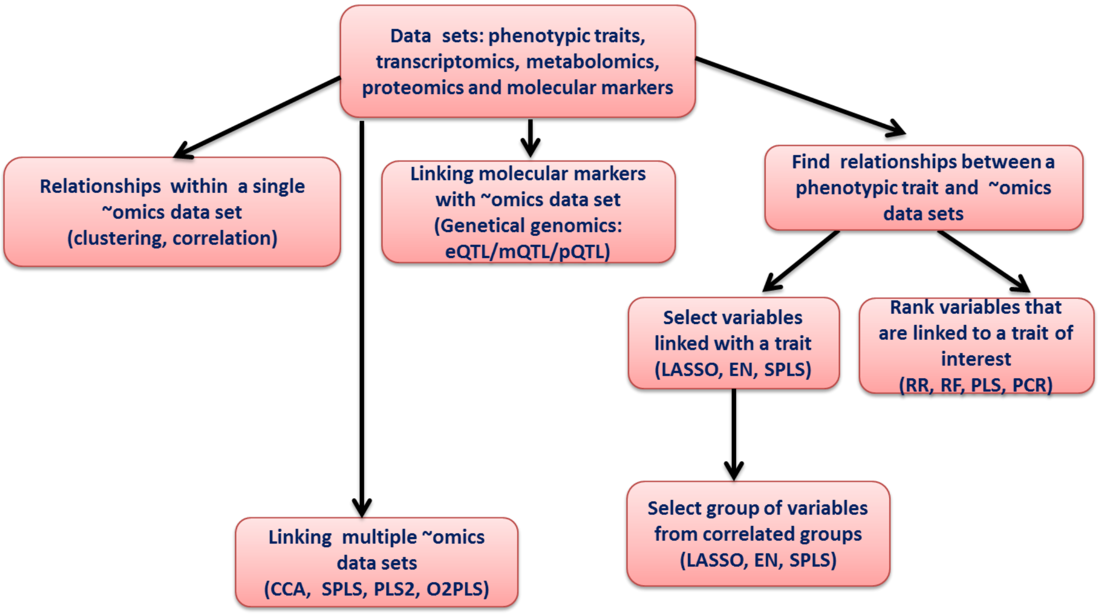

Data sets: phenotypic traits, transcriptomics, metabolomics, proteomics and molecular markers
Relationships within a single ~omics data set (clustering, correlation)
Linking molecular markers with ~omics data set (Genetical genomics: eQTL/mQTL/pQTL)
Find relationships between a phenotypic trait and ~omics data sets
Select variables linked with a trait (LASSO, EN, SPLS)
Rank variables that are linked to a trait of interest (RR, RF, PLS, PCR)
Linking multiple ~omics data sets (CCA, SPLS, PLS2, O2PLS)
Select group of variables from correlated groups (LASSO, EN, SPLS)

Figure 2: Diagram of the objectives given different ~omics data sets, marker data and phenotypic traits. Within brackets possible methods are listed. Elastic net = EN, principal components regression = PCR, partial least squares regression = PLS, sparse partial least squares regression = SPLS, Random Forest regression = RF, Canonical correlation analysis = CCA, Orthogonal partial least squares regression = O2PLS. Application of those methods depends the research question and the availability of the required data sets.

4 Integration of two (or more) ~omics data sets

When comparing multiple ~omics data sets, we are not dealing with a single response variable versus a multivariate ~omics data set, but rather two or more multivariate data sets for which we want to find relationships. In this situation we are also not predominantly interested in predicting one data set from another one (or multiple others), but we consider the relationships between the data sets in any direction. For example, we want to find relationships between a transcriptomics data set with a proteomics data set or any other type of ~omics data set with each other.

Integrating two data sets from different analytical platforms can enable an improved understanding of some underlying biological mechanisms (for example: LC-MS with GC-MS) and interactions between different functional levels, for example the gene expression and metabolite level. The advantage of this approach is that we can get integrated results from multiple correlated metabolites and genes at the same time**.** Different attempts have been made to integrate multiple ~omics data sets from different species such as metabolomics and proteomics in *Arabidopsis thaliana* using principal components analysis (PCA) and independent components analysis (ICA) (Wienkoop et al., 2008), and transcriptomics and metabolomics in *Arabidopsis thaliana* using orthogonal partial least squares regression (O2PLS) (Bylesjo et al., 2007), and transcriptomics, metabolomics and proteomics in grapevine berry also using O2PLS (Zamboni et al., 2010). A regularized canonical correlation analysis (rCCA) (Waaijenborg and Zwinderman 2009) used to link a gene-expression microarray and DNA-markers data set. PLS2 has been successfully applied to integration of gene expression and clinical data (with bridge PLS, Gidskehaug et al., 2007) to find a set of clinical traits correlated with gene expression.

In this thesis, we did not integrate multiple ~omics data sets irrespective of a trait of interest mainly because our main focus was not to discover possible links among different ~omics data sets, but, instead, to relate a phenotypic trait of interest (for example: tuber flesh colour of potato) to multiple ~omics data sets. As a consequence, we will miss direct links between metabolites, proteins and gene expression signals if they do not also have a connection to a phenotypic trait.

5 Genetical genomics studies in potato

The limitation of a direct study of the relationships between phenotypic traits and ~omics data is that they do not explicitly take the inheritance into account, so that it remains unclear whether a relationship is (partly) due to genetics or not. Therefore, we also performed QTL analyses of phenotypic traits and a genetical genomics study (Jansen and Nap 2001) in order to identify genetic loci that explain the variation in phenotypic traits, gene expression, primary and secondary metabolites and protein expression in potato tubers. In a genetical genomics approach, quantified gene expression levels are regarded as molecular quantitative traits that can be used in a QTL analysis. Similarly, metabolite and protein abundances can also be considered as molecular quantitative traits and analyzed by QTL analyses (Doerge 2002; Schadt et al., 2003). QTLs of gene expression profiles are denoted as expression QTLs (eQTL; Schadt et al., 2003). By analogy, QTLs for proteomics and metabolomics data are called protein QTLs (pQTLs) and metabolite QTLs (mQTLs), respectively (Keurentjes et al., 2006; Kliebenstein 2009; Acharjee et al., 2011; Keurentjes 2009; Matsuda et al., 2012)

In chapter 3 we performed mQTL analysis for LC-MS data in order to find genetic variation in the amounts of secondary metabolites. For example, we mapped QTLs for the putative metabolites 4,7-Megastigmadiene-3,9-diol-glucoside and 2,3-Dihydroxy-4-megastigmen-9-one-glucoside, which are non-volatile glucosides of carotenoid-derived volatile metabolites to chromosome 3 near the beta-carotene hydroxylase (Bch) gene which is one of the genes responsible for flesh colour in potato (Brown et al., 2003).

In chapter 4, mQTL analysis with GC-MS data was done to explain the variation of the primary metabolism in the CxE potato population. Further, we investigated associations between phenotypic and metabolic traits through QTL co-localization, correlation and Random Forest (RF) analyses. QTLs identify polymorphic loci that is contribute to the variation observed for a given trait (Causse et al., 2004). When QTLs for two different traits co-localize this might indicate the existence of a common regulator that controls the variation of both traits. However, co-localization of QTLs can also be due just to linkage of genes not necessarily involved in the same genetic pathway. In this thesis we performed Random Forest analyses and evaluated Pearson correlation coefficients to confirm associations between metabolites and phenotypes through mQTL studies. In addition, we have validated associations (using RF method on GC-MS data from cultivar data set)of putative metabolic peaks in a set of potato varieties and we confirm about the relationship between beta-alanine and starch phosphorylation. These two approaches: mQTL analysis and an association study between phenotypic traits and metabolites resulted in the determination of a strong relationship between potato starch phosphorylation and primary metabolism. Furthermore, our analyses resulted in identification of β-alanine as an important predictor for the degree of phosphorylation of starch in potato tubers. Inspection of the annotation of the metabolites selected after RF analysis included a number of unknown metabolites and more interestingly a few amino acids for which we also detected mQTLs coinciding with QTLs for phosphate content. Among these relevant metabolites, β-alanine was of particular interest because it consistently ranked in the top metabolites in different potato material used for the analyses. To explain the role of β-alanine in a biological sense is another matter and something for future research.

In chapter 5, we did protein QTL (pQTL) and phenotypic trait QTL analysis of quality traits in order to identify genetic loci that explain the variation in protein expression and variation in tuber quality traits. Further we studied co-localization of these QTLs and correlation of protein spots with quality traits. Here again, if we correlate phenotypic traits to proteomics data, we can find proteins that are related to the phenotype, but if we had only this information, the correlation could be independent of genetics and more related to environmental conditions influencing both the phenotypic trait and the protein abundance(s). Combining both QTL analysis and correlation analysis (also in chapter 4) gives us a better understanding about whether a genetic component is involved in this relationship and such a study could point to candidate proteins which are not only associated with the phenotype but which also show a *genetic* association.

5.1 Beyond QTL mapping

In this thesis, we used QTL analysis for ~omics data sets in a diploid potato mapping population (in chapter 3: eQTL and LC-MS mQTL, in chapter 4: for GC-MS mQTL, in chapter 5: pQTL) which includes only a small part of the existing allelic variation, which is one of the limitations of QTL analysis in a single cross population. This procedure is likely to identify only a fraction of the loci involved in the control of a trait. Further, QTL analysis is still problematic for species with complex polyploid genomes and this is quite common in many crop plants (i.e. cultivated potato is an autotetraploid). For studying the broad genetic architecture of complex traits, there is a need to be able to use germplasm covering a wider spectrum of alleles.

An association mapping (linkage disequilibrium mapping or genome wide association study) approach is well suited for this because it scrutinizes the results of many generations of recombination and selection (Syvanen 2005). Association mapping is increasingly being adopted as a genetic method complementary to traditional QTL mapping. The main advantages of association mapping are exploitation of allelic diversity from a collection of various more or less related cultivars and breeding materials. In addition, a higher mapping resolution may be reached as many more meiotic recombination events are sampled compared to a bi-parental segregating mapping population. Application of association mapping also has more advantages particularly in crops that are limited to no more than one generation per year (Flint-Garcia et al., 2003). Association mapping has been successfully applied for quality traits in tetraploid potato (D'hoop et al., 2009, 2010). It should also be noted that in an association mapping population, multiple alleles of the genes underlying traits can contribute to those traits. This often results in relatively minor effects exerted by many identified QTLs. Therefore, it is even more necessary to dissect complex traits into individual genetic parameters, and to use precision phenotyping of these traits for better analysis and understanding (Jackson et al., 2011). In such an association mapping panel similar studies could be undertaken as we have done here for the biparental C x E population but then with a subsequent better possible link to different alleles.

6 Targeted vs. untargeted analyses

In chapter 3 of this thesis we used metabolites, mainly carotenoids from 'targeted' analyses, which were extracted and analyzed by HPLC (*High-performance liquid chromatography*) with photodiode array (PDA) detection in addition to an untargeted metabolomics analysis from LC-MS analysis. These analyses included for example zeaxanthin and violaxanthin, compounds known to be linked to the carotenoid pathway (Brown et al., 2003) which is involved in tuber flesh colour of potato. The word "targeted" here refers to analysis of a small group of well-defined compounds that can be measured with techniques really adapted to the compounds of interest. For untargeted analyses, the emphasis is on analyzing a wide range of compounds with similar chemical properties (with respect to ionization, polarity, separability, solvability). Such untargeted analyses will measure a much larger number of compounds but often with more limited possibilities for direct identification or annotation. Such untargeted analyses include methods such as liquid chromatography-mass spectrometry (LC-MS), gas chromatography-mass spectrometry (GC-MS) and nuclear magnetic resonance (NMR). The change in analytical approaches from "targeted" to "untargeted" analyses has eliminated the limitation of searching only specific compounds of interest for which prior knowledge is needed. Through untargeted analysis it is now possible to generate a huge volume of data without specifying a priori any particular compounds of interest.

However, to uncover meaningful biological relationships (for example relationships between a phenotypic trait and a set of metabolites, as in chapter 2) from a large volume of data sets, there is a need for storage of the data through different databases (Mochida and Shinozaki 2010, 2011) and retrieve data and analyze them in a meaningful way, for example by relating phenotypes of interest to the available ~omics data sets using modern statistical methodology and bioinformatics infrastructure. It becomes crucial to develop statistical approaches or adapt existing ones based on the need of the research community, and to simplify and visualize results in an easy understandable way for researchers or plant breeders. In chapter 2, we developed a web application called "OmicsFusion" (http://www.plantbreeding.wur.nl/omicsFusion/) to link a trait of interest with an ~omics data set and summarize the results in a convenient way. Such a user interface is useful for the biologist and/or breeders to select and/or predict which genes (from gene expression data), metabolite or proteins have a connection to the trait of interest and hence this software tool can give a list of potential leads for further study, in terms of genes, proteins and/or metabolites.

7 Validation of the statistical results

An important aspect of ~omics studies as performed in the chapters of this thesis is validation or confirmation of the statistical results for example, in chapter 2 of this thesis, we related tuber flesh colour of potato to a metabolomics data set through different regression methods. As an outcome of the results we obtained two putative metabolic peaks which still need to be validated and identified by other experiments. 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 (Acharjee et al., 2011). Another example is present in chapter 3, where we validated results from transcriptomics analysis on the expression of genes such as bch (beta-carotene hydroxylase) and zep (Zeaxanthin epoxidase) by comparing our results to previous findings such as reported in Brown et al., 2006 and Wolters et al., 2010. In the same way, in chapter 4, we did GC-MS analysis aiming to find primary metabolites related to starch phosphorylation as our trait of interest. We found a relationship between β-alanine, an amino acid, to starch phosphorylation and validated this result by performing an independent experiment on potato cultivars because we did not find scientific literature about the relationship between β-alanine and starch phosphorylation. This might happen for two reasons: first, the link between the amino acid and starch phosphorylation is novel and second, it could be a false positive result. To avoid false positive results we generated GC-MS data in the cultivars and confirmed the evidence of β-alanine on potato cultivars using an association study with RF. In chapter 5, we did proteomics analysis of quality traits in potato tubers based on genetical genomics and univariate regression. In this case, we could not validate all our results because of the lack of annotation of the protein data and their identification. However, such a study can give a clue about candidate proteins and their possible involvement in a certain pathway. A QTL for the abundance of a protein, number 1129, was mapped to chromosome 3 at 80.8 cM and co-localizes with the flesh colour QTL and was also highly and positively correlated with flesh colour. From a previous study by Kloosterman et al., 2010, it was reported that carotenoids are involved in flesh colour and the *bch* gene plays a major role in flesh colour variation in potato. From this, we can hypothesize that protein number 1129 is involved in carotenoid biosynthesis which is responsible for flesh colour variation in potato.

In this thesis, almost in all the chapters we see the need for validation of our statistical analysis results. So, it is important to investigate or make a proper plan for validation which might involve follow-up experiments which can support statistical outcomes. Several studies such as Ioannidis and Khoury 2011 and Leek et al., 2012 also emphasized the need for proper experimental and / or independent validation of the ~omics research.

8 Marker- vs. ~omics-assisted breeding

Improvement of crops for quality and yield has been a fundamental question in plant breeding since cultivation began some ten thousand years ago. The development of molecular marker techniques in the early 1980s revolutionized plant breeding. Marker-assisted breeding (MAB) is a method that uses genetic markers that are linked to loci involved in a phenotypic trait of interest and that can be used in selection for that trait. In recent years there is an increasing interest in understanding natural variation in plants to study complex traits influenced by quantitative trait loci (QTL). Many of these studies focus on identification of QTLs underlying yield, product quality, and tolerance and resistance to abiotic and biotic stresses. However, recent metabolic and proteomic profiling is adding value in addition to QTL mapping in order to identify genetic loci that explain the variation in primary and secondary metabolites or protein expression in addition to those that explain variation in phenotypic traits.

Metabolite and protein levels are more closely linked to phenotype than genes (Fiehn 2002) and information about the metabolome or proteome of a given sample, for instance in metabolite profiling experiments can be obtained independent of genetic information and can lead to identification of metabolites that can be used as metabolic biomarkers (Steinfath et al., 2010; Carreno-Quintero et al., 2012; Fernie and Schauer 2009).

The use of metabolomics as a platform to find biomarkers for plant breeding (Meyer et al., 2007) has not yet been commercially exploited, because of the environmental and experimental variation, which can have a strong impact on metabolic profiles and also because limited annotation is still usually available of metabolic peaks in a metabolomics study (see also chapter 3). Thus, experimental design, sample preparation and annotation of the metabolic peaks are crucial parts of metabolomics studies and need to be performed carefully. On the other hand, in molecular marker applications, the DNA sequence is stable under any environmental condition, so that environmental and experimental conditions have hardly any influence. Despite the challenges of using metabolite biomarker discovery and application, recent studies in potato (Steinfath et al., 2010; Carreno-Quintero et al., 2012) suggest it has a potentially high value and opportunities for plant breeding using the predictive power of metabolites, for example in situations where for a trait under study the underlying biochemical mechanisms are still unknown.

9 Networks

In this thesis we used network analyses. Network analysis can give an integrative picture of the molecules (genes, proteins and metabolites) in terms of their interactions and regulation. Networks which involve only one layer of molecules (layer with only genes or metabolites or proteins) are called intra-level networks. Another type of networks involves multiple molecular levels and is called inter-level networks, like gene-metabolic networks involving genes and metabolites, gene-protein networks which involve genes and proteins, metabolite-protein networks with metabolite levels and proteins or gene-metabolite-protein networks, involving all three levels (in chapters 3 and 6 we used inter-level networks). In chapter 3, a network was made based on correlation coefficients among traits (flesh colour and enzymatic discoloration), gene expression and LC-MS metabolite peaks. The sizes of the correlations represent the strength of the interaction (edge) while the genes of which expression was measured, the metabolites peak and the phenotypic traits of interest represented the nodes in this network. In chapter 6, in the network analysis we tried to narrow down the number of genes, metabolites, proteins by selecting only one gene per chromosomal region involved and then performed a network analysis based on Pearson correlations between a trait, gene expression, GC-MS and LC-MS peaks, and protein expression.

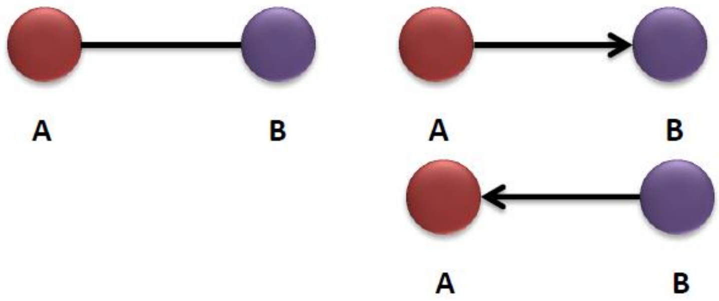


Figure 3: A gene for which gene expression was measured is shown as node A (in maroon) and a metabolite peak as node B (in purple). The left side Fig. shows that either A is activating/repressing B or the other way around. In the right side Fig., arrows represent directionality in either activation or repression of A by B, or of B by A.

One of the limitations of such correlation networks is that it does not say anything about the directionality of the regulation. For example in Fig. 3 (left side), given a gene expression from the gene at node A and metabolite abundance from metabolite B, in a correlation study we do not know whether A is activating/repressing B or the other way around or that activation /repression is even indirect, through another gene or metabolite. In other words, the use of correlation networks not only confounds direct and indirect associations but also provides no means to distinguish between cause and effect (Rhein and Strimmer 2007).

For "causal" analysis (Blair et al., 2012) typically the inference of a directed graphical model is required. Therefore, causal analysis requires tools and data different from correlation networks (e.g. time series data, motif information etc.) and this is out of the scope of this thesis.

10 Next generation sequencing and plant breeding

With the advent of next generation sequencing (NGS) technologies sequencing of RNA and DNA is also actively used in practical plant breeding (Gómez 2011). Application of NGS in maize and rice will lead to the widespread application in plant improvement by providing more markers and revealing candidate genes. In maize, high-resolution mapping using SNPs for flowering time, a quantitative trait, uncovered a large number of small-effect quantitative trait loci that acted in an additive fashion to determine flowering time (Buckler et al., 2009). In rice, resequencing and phenotyping allowed the identification of a large number of QTLs controlling 14 different agronomic traits (Huang et al., 2010). The genetic mapping of QTLs has been ongoing for many years. These QTLs tell us, with some statistical backup, the regions of chromosomes in which gene(s) affecting the trait are likely to reside. Ultimately, though, to fully take advantage of the QTLs, it is necessary to determine the identity of the genes responsible for the variation in the trait (Mackay 2001) and to understand the molecular basis of these QTLs (Hansen et al., 2008). The integration of genetic maps with associated QTLs and physical maps is a necessary aid to making these connections.

The availability of high-quality whole-genome sequence assemblies for major crops such as soybean (Schmutz et al., 2010), maize (Schnable et al., 2009) and potato (Xu et al., 2011) creates a paradigm change in how we can approach crop improvement. We now have access to all of the many thousands of genes that make up an organism. Resequencing of old varieties, landraces and even more newly released cultivars has the potential to uncover allelic diversity that has not been seen before, and to draw our attention to the regions of the genome that breeders have unknowingly focused upon in historical breeding efforts.

Now questions arise about which type of ~omics data are important to study a trait? From our study it is clear that how informative a particular ~omics data set is largely depends on the trait and biological information available for the trait of interest. If the trait is qualitative in nature, metabolomics data will be useful to investigate further the biochemical pathways and arise at potential candidate genes (for flesh colour; carotenoid pathway). But given any trait, gene expression data seems to be very useful anyway, to check the variation at the gene expression level linked with the trait. However, if we do not have any prior information about the trait of interest (novel trait), it will be difficult to say what type of ~omics experiments will be useful and give the highest chance to arrive at new candidate genes.

References

Acharjee A, Kloosterman B, de Vos RCH, Werij JS, Bachem CWB, Visser RGF, Maliepaard C (2011) Data integration and network reconstruction with omics data using Random Forest regression in potato. Analytica Chimica Acta, 705:56-63

Aebersold R and Mann M (2003) Mass spectrometry-based proteomics. Nature, 422: 198-207

Anithakumari AM, Tang J, van Eck HJ, Visser RGF, Leunissen JAM, Vosman B, van der Linden CG (2010) A pipeline for high throughput detection and mapping of SNPs from EST databases. Mol Breeding, 26:65–75

Bachem CWB, Hoeven RS, Bruijn SM, Vreugdenhil D, Zabeau M, Visser RGF (1996 Visualization of differentialgene expression using a novel method of RNA fingerprinting based on AFLP: analysis of gene expression during potato tuber development.The Plant Journal, 9:745-753

Bansal M, Della Gatta G, Bernardo D (2006) Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. Bioinformatics, 22: 815-822

Barabási A and Oltvai ZN (2004) Network biology: understanding the cell’s functional organization. Nature Reviews Genetics, 5:101-113

Barabási AL, Gulbahce N, Loscalzo J (2011) Network medicine: a network-based approach to human disease. Nature Reviews Genetics, 12:56-68

Batagelj V and Mrvar A (2003) Pajek - Analysis and Visualization of Large Networks Graph Drawing Software. Springer, Berlin

Benjamini Y and Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, 57, 289-300

Bernardo D, Thompson MJ, Gardner TS, Chobot SE, Eastwood EL, Wojtovich AP, Elliott SJ, Schaus SE, Collins JJ (2005) Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. Nature Biotechnology, 23:377 – 383

Biliaderis CG, Maurice TJ, Vose JR (1980) Starch Gelatinization Phenomena Studied by Differential Scanning Calorimetry. Journal of Food Science, 45:1669

Blair RH, Kliebenstein DJ, Churchill GA (2012) What Can Causal Networks Tell Us about Metabolic Pathways? PLoS Computational Biology, 8:1-12

Bonierbale MW, Plaisted RL, Tanksley SD (1988) Rflp Maps Based on a Common Set of Clones Reveal Modes of Chromosomal Evolution in Potato and Tomato, Genetics,120:1095-1103

Bormann CA, Rickert AM, Castillo Ruiz RA, Paal J, Lübeck J, Strahwald J, Buhr K, Gebhardt C (2004) Tagging Quantitative Trait Loci for Maturity-Corrected Late Blight Resistance in Tetraploid Potato with PCR-Based Candidate Gene Markers. Molecular Plant-Microbe Interactions, 17:1126-1138

Bøvelstad HM, Nygård S, Størvold HL, Aldrin M, Borgan Ø, Frigessi A, Lingjærde OC (2007) Predicting survival from microarray data - a comparative study. Bioinformatics, 23:2080-2087

Bøvelstad HM, Nygard S, Borgan Ø (2009) Survival prediction from clinico-genomic models - a comparative study. BMC Bioinformatics, 10, 413

Brazma A and Vilo J (2000) Gene expression data analysis. FEBS Journal, 480:17-24.

Breiman L (2001) Random forests. Machine Learning, 45:5-32.

Breithaupt DE and Bamedi A (2002) Carotenoids and carotenoid esters in potatoes(Solanum tuberosum L.): New insights into an ancient vegetable, Journal of Agricultural and Food Chemistry, 50:7175-7181

Breitling R, Li Y, Tesson BM, Fu J, Wu C, Wiltshire C, Gerrits A, Bystrykh LV, de Haan G, Su AI, Jansen RC (2008) Genetical Genomics: Spotlight on QTL Hot spots. PLoS Genetics, 4:1-4

Broman KW and Sen S (2009) A guide to QTL mapping with R/qtl, Springer

Brown CR, Kim TS, Ganga Z, Haynes K, De Jong D, Jahn M, Paran I, De Jong W (2006) Segregation of total carotenoid in high level potato germplasm and its relationship to beta-carotene hydroxylase polymorphism. American Journal of Potato Research, 83:365-372

Brown CR, Wrolstad R, Durst R, Yang CP,Clevidence B(2003) Breeding studies in potatoes containing high concentrations of anthocyanins, American Journal of Potato Research, 80:241-249

Brown CR, Edwards CG, Yang CP, Dean BB (1993) Orange Flesh Trait in Potato - Inheritance and Carotenoid Content, Journal of the American Society for Horticultural Science, 118:145-150

Bryan GT, Wu KS, Farrall L, Jia Y, Hershey HP, McAdams SA, Faulk KN, Donaldson GK, Tarchini R, Valent B (2000) A single amino acid difference distinguishes resistant and susceptible alleles of the rice blast resistance gene Pi-ta. Plant Cell, 12:2033-2045

Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ, Browne C, Ersoz E, Flint-Garcia S, Garcia A, Glaubitz JC, Goodman MM, Harjes C, Guill K, Kroon DE,Larsson S, Lepak NK, Li H, Mitchell SE, Pressoir G, Peiffer JA, Rosas MO, Rocheford TR, Romay MC, Romero S,Salvo S, Villeda HS, Silva HSD, Sun Q, Tian F, Upadyayula N, Ware D, Yates H, Yu J, Zhang Z, Kresovich S, McMullen MD (2009) The genetic architecture of maize flowering time. Science, 325: 714-718

Bylesjo M, Eriksson D, Kusano M, Moritz T, Trygg J (2007) Data integration in plant biology: the O2PLS method for combined modeling of transcript and metabolite data. The Plant Journal, 52:1181-1191

Camacho D, Fuente A, Mendes P (2005) The origin of correlations in metabolomics data Metabolomics, 1:53-63

Campbell R, Ducreux LJM, Morris W L, Morris JA, Suttle JC, Ramsay G, Bryan GJ, Hedley PE, Taylor MA (2010) The metabolic and developmental roles of carotenoid cleavage dioxygenase4 from potato. Plant Physiol, 154: 656-664

Carreno-Quintero N, Acharjee A, Maliepaard C, Bachem CW, Mumm R, Bouwmeester H, Visser RG, Keurentjes JJ (2012) Untargeted metabolic quantitative trait loci analyses reveal a relationship between primary metabolism and potato tuber quality.Plant Physiol,158:1306-18

Caspar T, Lin T-P, Kakefuda G, Benbow L, Preiss J, Somerville C (1991) Mutants of Arabidopsis with Altered Regulation of Starch Degradation. Plant Physiology, 95: 1181-1188

Caswell JA and Mojduszka EM (1996) Using informational labeling to influence the market for quality in food products, American Journal for Agricultural Economics 78:1248-1253

Causse M, Duffe P, Gomez MC, Buret M, Damidaux R, Zamir D, Gur A, Chevalier C, Lemaire-Chamley M, Rothan C (2004) A genetic map of candidate genes and QTLs involved in tomato fruit size and composition. Journal of Experimental Botany, 55:1671-1685

Celis-Gamboa C, Struik P, Jacobsen E, Visser RGF (2003) Temporal dynamics of tuber formation and related processes in a crossing population of potato (Solanum tuberosum). Annals of Applied Biology, 143:175-187

Celis-Gamboa BC (2002) The life cycle of the potato (Solanum tuberosum L.): from crop physiology to genetics, Ph.D. Thesis, Wageningen University, Wageningen, The Netherlands, 191pp, ISBN 90-5808-688-7

Chakauya E, Coxon KM, Whitney HM, Ashurst JL, Abell C, Smith AG (2006) Pantothenate biosynthesis in higher plants: advances and challenges. Physiologia Plantarum, 126: 319-329

Chan EKF, Rowe HC, Kliebenstein DJ (2010) Understanding the Evolution of Defense Metabolites in Arabidopsis thaliana Using Genome-wide Association Mapping. Genetics, 185:991-1007

Chen X, Hackett CA, Niks RE, Hedley PE, Booth C, Druka A, Marcel TC, Vels A, Bayer M, Milne I, Morris J, Ramsay L, Marshall D, Cardle L and Waugh R (2010) An eQTL analysis of partial resistance to Puccinia hordei in barley. PLoS One. 5(1):e8598.

Chun H and Keles S (2009) Expression quantitative trait loci mapping with multivariate sparse partial least squares regression. Genetics, 182:79-90

Cheverud JM (2001) A simple correction for multiple comparisons in interval mapping genome scans. Heredity, 87: 52–58

Coffin RH, Yada RY, Parkin KL, Grodzinski B, Stanley DW (1987) Effect of Low Temperature Storage on Sugar Concentrations and Chip Color of Certain Processing Potato Cultivars and Selections. Journal of Food Science, 52: 639-645

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

Collins A, Milbourne D, Ramsay L, Meyer R, Chatot-Balandras C, Oberhagemann P, De Jong W, Gebhardt C, Bonnel E, Waugh R (1999) QTL for field resistance to late blight in potato are strongly correlated with maturity and vigour. Molecular Breeding, 5: 387-398

Corsini D, Pavek J, Dean B (1992) Differences in free and protein-bound tyrosine among potato genotypes and the relationship to internal blackspot resistance. American Journal of Potato Research, 69:423-435

Cristianini N and Shawe-Taylor J (2000) An introduction to support vector machines. Cambridge Universtity Press

Dancs G, Kondrak M, Banfalvi Z (2008) The effects of enhanced methionine synthesis on amino acid and anthocyanin content of potato tubers. BMC Plant Biol, 8: 65

Davies HV (2007) Metabolomics: Applications in functional biodiversity analysis in potato. In Acta Hort, 745: 471-484

Davies HV, Shepherd LVT, Burrell MM, Carrari F, Urbanczyk-Wochniak E, Leisse A, Hancock RD, Taylor M, Viola R, Ross H, McRae D, Willmitzer L, Fernie AR (2005) Modulation of fructokinase activity of potato (Solanum tuberosum) results in substantial shifts in tuber metabolism. Plant and Cell Physiology, 46:1103-1115

Delmotte N, Ahrens CH, Knief C, Qeli E, Koch M, Fischer HM, Vorholt JA, Hennecke H, Pessi G (2010) An integrated proteomics and transcriptomics reference data set provides new insights into the Bradyrhizobium japonicum bacteroid metabolism in soybean root nodules. Proteomics, 10:1391-1400

Demiriz A, Bennett KP, Breneman CM, Embrechts MJ (2001). Support Vector Machine Regression in Chemometrics. Comp Sci Stat, 33:289-296

De Vos RCH, Moco S, Lommen A, Keurentjes JJ, Bino RJ, Hall RD (2007) Untargeted large-scale metabolomics using liquid chromatography coupled to mass spectrometry. Nat. Protoc, 2:778-791

D’hoop B (2009) Association mapping in tetraploid potato, PhD Thesis Wageningen University, The Netherlands

D'Hoop BB, Paulo MJ, Kowitwanich K, Sengers M, Visser RGF, Van Eck HJ, Van Eeuwijk F A (2010) Population structure and linkage disequilibrium unraveled in tetraploid potato. Theor Appl Genet,121:1151-1170

D’hoop BB, Paulo JM, Mank RA, Eck HJV, Eeuwijk FAV (2008) Association mapping of quality traits in potato (Solanum tuberosum L.). Euphytica, 161:47-60

Dietterich TG (1999) Machine Learning. The MIT Encyclopedia of the Cognitive Sciences, MIT Press

Di R, Kim J, Martin MN, Leustek T, Jhoo J, Ho C-T, Tumer NE (2003) Enhancement of the Primary Flavor Compound Methional in Potato by Increasing the Level of Soluble Methionine. Journal of Agricultural and Food Chemistry, 51: 5695-5702

Dobson G, Shepherd T, Verrall SR, Conner S, McNicol JW, Ramsay G, Shepherd LVT, Davies HV, Stewart D (2008) Phytochemical Diversity in Tubers of Potato Cultivars and Landraces Using a GC-MS Metabolomics Approach. Journal of Agricultural and Food Chemistry, 56: 10280-10291

Dobson G, Shepherd T, Verrall SR, Griffiths WD, Ramsay G, McNicol JW, Davies HV, Stewart D (2009) A Metabolomics Study of Cultivated Potato (Solanum tuberosum) Groups Andigena, Phureja, Stenotomum, and Tuberosum Using Gas Chromatography−Mass Spectrometry. Journal of Agricultural and Food Chemistry, 58: 1214-1223

Doerge RW (2002) Mapping and analysis of quantitative trait loci in experimental populations. Nature Review Genetics, 3:43-52

Dunn WB, Bailey NJC, Johnson HE (2005). Measuring the metabolome: current analytical technologies. Analyst, 130: 606-625

Díaz-Uriarte R and Alvarez de Andrés S (2006) Gene selection and classification of microarray data using random forest. BMC Bioinformatics, 7: 3

Fernie AR and Schauer N (2009) Metabolomics-assisted breeding: a viable option for crop improvement? Trends in genetics, 25:39-48

Fiehn O (2002). Metabolomics - the link between genotypes and phenotypes. Plant Molecular Biology, 48:155-171

Flint-Garcia SA, Thornsberry JM, S E, IV B (2003) Structure of linkage disequilibrium in plants. Ann Rev Plant Biol, 54:357-374

Freedman DA (2005) Statistical Models: Theory and Practice Cambridge, UK: Cambridge University Press

Friedman J, Hastie T, Tibshirani R (2010) A note on the group lasso and a sparse group lasso, technical report

Fu J, Swertz MA, Keurentjes JJB, Jansen RC (2007) MetaNetwork: a computational protocol for the genetic study of metabolic networks. Nat. Protocols, 2: 685-694

Fukushima A, Kusano M, Redestig H, Arita M, Saito K (2009) Integrated omics approaches in plant systems biology. Current Opinion in Chemichal Biology, 13:532-538

Gaasterland T and Bekiranov S (2000) Making the most of microarray data. Nature Genetics, 24: 204-206

Geladi P and Kowlaski B (1986) Partial least square regression: A tutorial. Analytica Chimica Acta, 35:1-17

Ghosh S (2008) Regularization as a Toolkit for Parsimonious Modeling in Bioinformatics, preprint

Gidskehaug L, Anderssen E, Flatberg A , Alsberg BK (2007) A framework for significance analysis of gene expression data using dimension reduction methods. BMC Bioinformatics, 8:346

Gislason PO, Benediktsson JA, Sveinsson JR (2006) Random forest for land cover classification. Pattern Recognition Letters,4:294-300

Gómez JMJ (2011) Next generation quantitative genetics in plants. Frontiers in Plant science, 2:1-10

Gonzalez I and Dejean S (2008) CCA: An R Package to Extend Canonical Correlation Analysis, Journal of Statistical Software , 23

Griffiths WJ and Wang Y (2009) Mass spectrometry: from proteomics to metabolomics and lipidomics. Chemical Society Reviews, 38:1882-1896

Han JDJ (2008) Understanding biological functions through molecular networks cell research, 18:224-237

HastieT, Tibshirani R, Friedman J (2001) The elements of statistical learning: data mining, inference, and prediction. Springer-Verlag, New York

Hannemann RE and Peloquin SJ (1967) Crossability of 24-chromosome potato hybrids with 48-chromosome cultivars. Potato Research,10:62-73

Halford NG, Hey S, Jhurreea D, Laurie S, McKibbin RS, Zhang Y, Paul MJ (2004) Highly conserved protein kinases involved in the regulation of carbon and amino acid metabolism. Journal of Experimental Botany, 55:35-42

Hansen BG, Halkier BA, Kleibenstein DA (2008) Identifying the molecular basis of QTLs: eQTLs add a new dimension. Trends in Plant Science, 13:1360-1385

Haley CS and Knott SA (1992) A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity, 69:315-324

Heffner EL, Sorrells ME, Jannink JL(2009) Genomic selection for crop improvement. Crop Science, 49:1-12

Heslot N, Yang HP, Sorrells ME, Jannink JL(2012) Genomic Selection in Plant Breeding: A Comparison of Models. Crop Science, 52:146-160

Hendriks MM, Smit S, Akkermans WL, Reijmers TH, Eilers PH, Hoefsloot HC, Rubingh CM, de Koster CG, Aerts JM, Smilde AK (2007). How to distinguish healthy from diseased? Classification strategy for mass spectrometry-based clinical proteomics. Proteomics, 7:3672-3680

Hoskuldson A (1988) PLS regression methods. Journal of Chemometrics, 2:211-228

Hoerl AE and Kennard RW (1970) Ridge regression: biased estimation for non-orthogonal problems. Technometrics, 12:55-67

Hizukuri S, Tabata S, Kagoshima, Nikuni Z (1970) Studies on Starch Phosphate Part 1. Estimation of glucose-6-phosphate residues in starch and the presence of other bound phosphate(s). Starch – Stärke, 22: 338-343

Huang X, Wei X, Sang T, Zhao Q, Feng Q, Zhao Y, Li C, Zhu C, Lu T, Zhang Z, Meng Li, Fan D,Guo Y, Wang A,Wang L, Deng L, Li W,Lu Y, Weng Q, Liu K,Huang T, Zhou T, Jing Y, Li W, Zhang Lin,Buckler ES, Qian Q, Zhang QF,Li J, Han B(2010) Genome-wide association studies of 14 agronomic traits in rice landraces. Nature Genetics, 42:961-967

Hurtado PX (2012) Investigating genotype by environment and QTL by environment interactions for developmental traits in potato,Ph.D. Thesis, Wageningen University, Wageningen, The Netherlands, 65pp ISBN 978-94-6173-438-9

Ioannidis JPA and Khoury MJ (2011) Improving Validation Practices in "Omics" Research. Science, 334:1230-1232

Jacobsen E (1980) Increase of diplandroid formation and seed set in 4x X 2x crosses in potatoes by genetical manipulation of diphaploids and some theoretical consequences. Z Pflanzenzuecht, 85:10-121

Jacobs JME, Eck HJ, Arens P, Verkerk-Bakker B, te Lintel Hekkert B, Bastiaanssen HJM, El-Kharbotly A, Pereira A, Jacobsen E, Stiekema WJ (1995) A genetic map of potato (Solanum tuberosum) integrating molecular markers, including transposons, and classical markers. Theor Appl Genet., 91:289-300.

Jackson SA, Iwata A, Lee SH, Schmutz J, Shoemaker R (2011) Sequencing crop genomes: approaches and applications New Phytologist, 191: 915-925

Jansen R and Nap J (2001) Genetical genomics: the added value from segregation. Trends Genet, 17: 388-391

Jannink JL, Lorenz AJ, Iwata H (2010) Genomic selection in plant breeding:from theory to practice. Briefings in Fuctional Genomics, 9:166-177

Jolliffe IT (1982) A note on the use of principal components in regression. Journal of the Royal Statistical Society, 31:300-303

Joyce AR and Palsson B (2006) The model organism as a system: integrating 'omics' data sets. Nat. Rev. Mol. Cell Biol, 7:198-210

Jiang R, Tang W , Wu X, Fu W (2009) A random forest approach to the detection of epistatic interactions in case-control studies.BMC Bioinformatics 10 :S65

Kadam SS, Dhumal SS, Jambhale ND (1991) Structure, nutritional composition, and quality. Potato: Production,processing, and products. Boca Raton, Fla.: CRC Press: 9-36

Kearsey MJ(1998) The principles of QTL analysis (a minimal mathematics approach), Journal of Experimental Botany, 49:1619-1623

Kearsey MJ and Farquhar (1998) QTL analysis in plants; where are we now? Heredity, 80: 137-142

Keurentjes JJB (2009) Genetical metabolomics: closing in on phenotypes. Current Opinion in Plant Biology, 12:223-230

Keurentjes JJB, Fu JY, de Vos CHR, Lommen A, Hall RD, Bino RJ, van der Plas LHW, Jansen RC, Vreugdenhil D, Koornneef M (2006) The genetics of plant metabolism. Nature Genetics, 38:842-849

Kitano H (2002) Systems Biology: A Brief Overview. Science, 295:1662-1664

Kiers HAL and Smilde AK (2007) A comparison of various methods for multivariate regression with highly collinear variable. Statistical Methods and Applications, 16:193–228

Kim TY, Kim HU, Lee SY (2010) Data integration and analysis of biological networks. Curr. Opin. Biotech, 21:78-84

Kim YS, Wiesenborn DP, Grant LA (1997) Pasting and thermal properties of potato and bean starches. Starch-Starke, 49:97-102

Kleinkauf H (2000) The role of 4'-phosphopantetheine in the biosynthesis of fatty acids, polyketides and peptides. BioFactors, 11:91-92

Kliebenstein DJ (2007) Metabolomics and plant quantitative trait locus analysis-The optimum genetical genomics platform? In: Nikolau BJ, Wurtele ES, editors. Concepts in plant metabolomics. Dordrecht (the Netherlands): Springer. pp. 29-45

Kliebenstein D (2009) Quantitative Genomics: Analyzing Intraspecific Variation Using Global Gene Expression Polymorphisms or eQTLs. The Annual Review of Plant Biology, 60:93-114

Kloosterman B, Oortwijn M, uitdeWilligen J, America T, de Vos R, Visser RG, Bachem CW (2010) From QTL to candidate gene: genetical genomics of simple and complex traits in potato using a pooling strategy. BMC Genomics, 11:158

Kloosterman B, Anithakumari AM, Chibon PY, Oortwijn M, van der Linden GC, Visser RGF, Bachem CWB (2012) Organ specificity and transcriptional control of metabolic routes revealed by expression QTL profiling of source–sink tissues in a segregating potato population. BMC Plant Biology, 12:17

Kohyama K and Sasaki T(2006) Differential scanning calorimetry and a model calculation of starches annealed at 20 and 50 °C. Carbohydrate Polymers, 63:82-88

Lacroix V, Cottret L,Thébault P, Sagot MF (2008) An Introduction to Metabolic Networks and their Structural Analysis. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 5: 594-617

Lê Cao KA, Rossouw D, Robert-Granié C, Besse P(2008) A sparse PLS for variable selection when integrating Omics data. Stat. Appl. Genet. Mol. Biol., 7:1544-6115

Lê Cao KA, Martin PGP, Robert-Granié C and Besse P (2009) Sparse canonical methods for biological data integration: application to a cross-platform study. BMC Bioinformatics, 10:34

Leek JT, Taub MA, Rasgon JL (2012) A statistical approach to selecting and confirming validation targets in –omics experiments. BMC Bioinformatics, 13:150

Liu BH (2002) Statistical Genomics: Linkage, Mapping and QTL Analysis. 2nd Edition. CRC Press, Boca Raton, Florida

Lisec J, Meyer RC, Steinfath M, Redestig H, Becher M, Witucka-Wall H, Fiehn O, Torjek O, Selbig J, Altmann T, Willmitzer L (2008) Identification of metabolic and biomass QTL in Arabidopsis thaliana in a parallel analysis of RIL and IL populations. Plant Journal, 53: 960-972

Lisec J, Schauer N, Kopka J, Willmitzer L, Fernie AR (2006) Gas chromatography mass spectrometry-based metabolite profiling in plants. Nature Protocols,1:387-396

Li J and Ji L (2005) Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. Heredity, 95:221-227

Li XQ, De Jong H, De Jong DM, De Jong WS (2005) Inheritance and genetic mapping of tuber eye depth in cultivated diploid potatoes. Theoretical and Applied Genetics, 110:1068-1073

Li L, Paulo MJ, Strahwald J, Lubeck J, Hofferbert HR, Tacke E, Junghans H, Wunder J, Draffehn A, Eeuwijk AFV, Gebhardt C (2008) Natural DNA variation at candidate loci is associated with potato chip colour, tuber starch content, yield and starch yield. Theoretical and Applied Genetics, 116:1167-1181

Lommen A (2009) MetAlign: Interface-Driven, Versatile Metabolomics Tool for Hyphenated Full-Scan Mass Spectrometry Data Preprocessing. Analytical Chemistry, 81:3079-3086

Lorberth R, Ritte G, Willmitzer L, Kossmann J (1998) Inhibition of a starch-granule-bound protein leads to modified starch and repression of cold sweetening. Nature Biotechnology, 16:473-477

Mackay TFC (2001) The genetic architecture of quantitative traits. Annual Review of Genetics, 35:303-339

Massy WF (1965) Principal Components Regression in Exploratory Statistical Research. Journal of the Royal Statistical Society, 60:234-246

Markowetz F and Spang R(2007) Inferring cellular networks – a review. BMC Bioinformatics, 8

Masouleh AK, Waters DLE, Reinke RF, Henry RJ (2009) A high-throughput assay for rapid and simultaneous analysis of perfect markers for important quality and agronomic traits in rice using multiplexed MALDI-TOF mass spectrometry. Plant Biotechnology Journal, 7:355-363

Matsuda F, Okazaki Y, Oikawa A, Kusano M, Nakabayashi R, Kikuchi J, Yonemaru JI, Ebana K, Yano M, Saito K (2012) Dissection of genotype–phenotype associations in rice grains using metabolome quantitative trait loci analysis. The Plant Journal, 70:624-636

Maccaferri M, Sanguineti MC, Demontis A, El-Ahmed A, Garcia del Moral L, Maalouf F, Nachit M, Nserallah N, Ouabbou H, Rhouma S, Royo C, Villegas D, Tuberosa R (2011) Association mapping in durum wheat grown across a broad range of water regimes. J Exp Bot.,2:409-438

Massy WF (1965) Principal Components Regression in Exploratory Statistical Research. Journal of the Royal Statistical Society, 60:234-246

Mevik BH and Cederkvis HR (2004) Mean squared error of prediction (MSEP) estimates for principal component regression (PCR) and partial least squares regression (PLSR). Journal of Chemometrics, 18:422-429

Menendez CM, Ritter E, Scha¨fer-Pregl R, Walkemeier B, Kalde A, Salamini F, Gebhardt C (2002) Cold-sweetening in diploid potato. Mapping QTL and candidate genes. Genetics, 162:1423-1434

Meuwissen THE, Hayes BJ,Goddard ME(2001) Prediction of total genetic value using Genome-Wide dense marker maps. Genetics, 157:1819-1829

Meyer RC, Steinfath M, Lisec J, Becher M, Witucka-Wall H, Torjek O, Fiehn O, Eckardt A, Willmitzer L, Selbig J, Altmann T (2007) The metabolic signature related to high plant growth rate in Arabidopsis thaliana. Proceedings of the National Academy of Sciences of the United States of America, 104:4759-4764

Mikkelsen R, Baunsgaard L, Blennow A (2004) Functional characterization of alpha-glucan,water dikinase, the starch phosphorylating enzyme. Biochemical Journal, 377:525-532

Mondy NI and Munshi CB (1993) Effect of maturity and storage on ascorbic acid and tyrosine concentrations and enzymic discoloration of potatoes. Journal of Agricultural and Food Chemistry, 41:1868-1871

Morreel K, Goeminne G, Storme V, Sterck L, Ralph J, Coppieters W, Breyne P, Steenackers M, Georges M, Messens E, Boerjan W (2006) Genetical metabolomics of flavonoid biosynthesis in Populus: a case study. The Plant Journal, 47:224-237

Montgomery CD and Peck EA (1992). Introduction to Linear Regression Analysis, Wiley, New York

Mochida K and Shinozaki K (2010) Genomics and Bioinformatics Resources for Crop Improvement. Plant Cell Physiology, 51:497-523

Mochida K and Shinozaki K (2011) Advances in Omics and Bioinformatics Tools for Systems Analyses of Plant Functions. Plant Cell Physiology, 12:2017-2038

Moco S, Bino RJ, Vorst O, Verhoeven HA, de Groot J, Van Beek TA, Vervoort J, De Vos CH (2006) A Liquid Chromatography-Mass Spectrometry-Based Metabolome Database for Tomato.Plant Physiol., 141:1205-1218

Nikiforova VJ, Daub CO, Hesse H, Willmitzer L, Hoefgen R (2005) Integrative gene-metabolite network with implemented causality deciphers informational fluxes of sulphur stress response, Journal of Experimental Botany, 56:1887-96

Noctor G, Novitskaya L, Lea PJ, Foyer CH (2002) Coordination of leaf minor amino acid contents in crop species: significance and interpretation. Journal of Experimental Botany, 53: 939-945

Oberhagemann P, Chatot-Balandras C, Schäfer-Pregl R, Wegener D, Palomino C, Salamini F, Bonnel E, Gebhardt C (1999) A genetic analysis of quantitative resistance to late blight in potato: towards marker-assisted selection. Molecular Breeding, 5:399-415

Ozsolak F and Milos PM (2011) RNA sequencing: advances, challenges and opportunities. Nature Reviews Genetics, 12:87-98

Overy SA, Walker HJ, Malone S, Howard TP, Baxter CJ, Sweetlove LJ, Hill SA, Quick WP (2005) Application of metabolite profiling to the identification of traits in a population of tomato introgression lines. Journal of Experimental Botany, 56:287-296

Ogutu JO, Schulz-Streeck T, Piepho HP (2012) Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. BMC Proceedings, 2:S10

Patersona H, Lander ES, Hewitt JD, Petersons S, Lincoln E, Tanskley SD (1988) Resolution of quantitative traits into Mendelian factors, using a complete linkage map of restriction fragment length polymorphisms. Nature, 335:721-726

Park T and Casella G (2008) The Bayesian Lasso. Journal of the American Statistical Association, 103:681-686

Patterson SD and Aebersold R (2003) Proteomics: the first decade and beyond. Nature Genetics, 33:311-323.

Pang H, Lin A, Holford M, Enerson BE, Lu B, Lawton MP, Floyd E, Zhao H (2006) Pathway analysis using random forests classification and regression. Bioinformatics, 16:2028-2036

Prat S, Frommer WB, Höfgen R, Keil M, Kossmann J, Köster-Töpfer M, Liu XJ, Müller B, Peña-Cortés M, Rocha-Sosa M, Sánchez-Serrano JJ, Sonnewald U, Willmitzer L (1990) Gene expression during tuber development in potato plants. FEBS Letters, 286:334-338

Potato genome sequence consortium (2011) Genome sequence and analysis of the tuber crop potato. Nature, 475:189-9

Quackenbush J (2001) Computational analysis of microarray data. Nature review genetics, 2: 418

Rhein RO and Strimmer K (2007) From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. BMC Systems Biology, 1:37

Ritte G, Lloyd JR, Eckermann N, Rottmann A, Kossmann J, Steup M (2002) The starch-related R1 protein is an alpha-glucan, water dikinase. Proceedings of the National Academy of Sciences of the United States of America, 99:7166-7171

Roessner U, Wagner C, Kopka J, Trethewey RN, Willmitzer L (2000) Simultaneous analysis of metabolites in potato tuber by gas chromatography-mass spectrometry. Plant Journal, 23:131-142

Roessner U, Luedemann A, Brust D, Fiehn O, Linke T, Willmitzer L, Fernie AR (2001) Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. Plant Cell, 13:11-29

Roessner U, Wagner C, Kopka J, Trethewey RN, Willmitzer L (2000) Simultaneous analysis of metabolites in potato tuber by gas chromatography-mass spectrometry. Plant Journal, 23: 131-142

Rooke HS, Lampitt LH, Jackson EM (1949) The phosphorus compounds of wheat starch. Biochemical Journal, 45:231-236

Rowe HC, Hansen BG, Halkier BA, Kliebenstein DJ (2008) Biochemical networks and epistasis shape the Arabidopsis thaliana metabolome. Plant Cell, 20:1199-1216

Russo G, Bernardo M, Sontag ED (2010) Global entrainment of transcriptional systems to periodic inputs. PLoS Computational Biology, 6:1000739

Ruiz D and Egea J (2008) Phenotypic diversity and relationships of fruit quality traits in apricot (Prunus armeniaca L.) germplasm. Euphytica, 163:143-158

Rigal D, Gauillard F, Forget FR (2000) Changes in the carotenoid content of apricot (Prunus armeniaca, var Bergeron) during enzymatic browning: β-carotene inhibition of chlorogenic acid degradation. J Sci Food Agric., 80:763-768

Saeys Y, Inza I, Larranaga P (2007) A review of feature selection techniques in bioinformatics. Bioinformatics, 23: 2507-2517

Scott GJBR, Rosegrant M, Bokanga M(2000) Roots and tubers in the global food system: a vision statement to the year 2020, Lima (Peru)

Segal MR (2004) Machine Learning Benchmarks and Random Forest Regression. Technical Report, Center for Bioinformatics & Molecular Biostatistics, University of California, San Francisco

Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, Ruff TG, Milligan SB, Lamb JR, Cavet G, Linsley PS, Mao M, Stoughton RB, Friend SH (2003) Genetics of gene expression surveyed in maize, mouse and man. Nature, 422:297-302

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, Xu D, Hellsten U, May GD, Yu Y, Sakurai T, Umezawa T, Bhattacharyya MK, Sandhu D, Valliyodan B, Lindquist E, Peto M, Grant D, Shu S, Goodstein D, Barry K, Futrell-Griggs M, Abernathy B, Du J, Tian Z, Zhu L, Gill N, Joshi T, Libault M, Sethuraman A, Zhang XC, Shinozaki K, Nguyen HT, Wing RA, Cregan P, Specht J, Grimwood J, Rokhsar D, Stacey G, Shoemaker RC, Jackson SA (2010) Genome sequence of the palaeopolyploid soybean. Nature, 463:178-183

Schnable P, Ware D, Fulton R, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, Minx P, Reily AD, Courtney L, Kruchowski SS, Tomlinson C, Strong C, Delehaunty K, Fronick C, Courtney B, Rock SM, Belter E, Du F, Kim K, Abbott RM, Cotton M, Levy A, Marchetto P, Ochoa K, Jackson SM, Gillam B, Chen W, Yan L, Higginbotham J, Cardenas M, Waligorski J, Applebaum E, Phelps L, Falcone J, Kanchi K, Thane T, Scimone A, Thane N, Henke J, Wang T, Ruppert J, Shah N, Rotter K, Hodges J, Ingenthron E, Cordes M, Kohlberg S, Sgro J, Delgado B, Mead K, Chinwalla A, Leonard S, Crouse K, Collura K, Kudrna D, Currie J, He R, Angelova A, Rajasekar S, Mueller T, Lomeli R, Scara G, Ko A, Delaney K, Wissotski M, Lopez G, Campos D, Braidotti M, Ashley E, Golser W, Kim H, Lee S, Lin J, Dujmic Z, Kim W, Talag J, Zuccolo A, Fan C, Sebastian A, Kramer M, Spiegel L, Nascimento L, Zutavern T, Miller B, Ambroise C, Muller S, Spooner W, Narechania A, Ren L, Wei S, Kumari S, Faga B, Levy MJ, McMahan L, Van Buren P, Vaughn MW, Ying K, Yeh CT, Emrich SJ, Jia Y, Kalyanaraman A, Hsia AP, Barbazuk WB, Baucom RS, Brutnell TP, Carpita NC, Chaparro C, Chia JM, Deragon JM, Estill JC, Fu Y, Jeddeloh JA, Han Y, Lee H, Li P, Lisch DR, Liu S, Liu Z, Nagel DH, McCann MC, SanMiguel P, Myers AM, Nettleton D, Nguyen J, Penning BW, Ponnala L, Schneider KL, Schwartz DC, Sharma A, Soderlund C, Springer NM, Sun Q, Wang H, Waterman M, Westerman R, Wolfgruber TK, Yang L, Yu Y, Zhang L, Zhou S, Zhu Q, Bennetzen JL, Dawe RK, Jiang J, Jiang N, Presting GG, Wessler SR, Aluru S, Martienssen RA, Clifton SW, McCombie WR, Wing RA, Wilson RK. (2009)The b73 maize genome: complexity, diversity, and dynamics. Science, 326:1112–1115

Schauer N, Semel Y, Roessner U, Gur A, Balbo I, Carrari F, Pleban T, Perez-Melis A, Bruedigam C, Kopka J, Willmitzer L, Zamir D, Fernie AR (2006) Comprehensive metabolic profiling and phenotyping of interspecific introgression lines for tomato improvement. Nat Biotechnol., 24:447-454

Schachter RD and Kenley CR (1989) Gaussian influence diagrams. Management Sci, 35:527-550

Schlotterer C (2004) The evolution of molecular markers—just a matter of fashion? Nature Reviews Genetics, 5:63-69

Schafer-Pregl R, Ritter E, Concilio L, Hesselbach J, Lovatti L, Walkemeier B, Thelen H, Salamini F, Gebhardt C (1998) Analysis of quantitative trait loci (QTL) and quantitative trait alleles (QTA) for potato tuber yield and starch content. Theoretical and Applied Genetics, 97:834-846

Smith AM, Zeeman SC, Smith SM (2005) Starch degradation. Annual Review of Plant Biology, 56:73-98

Smit S, Breemen M, Hoefsloot HCJ, Smilde AK, Aerts JMFG, Koster CG (2007). Assessing the statistical validity of proteomics based biomarkers. Analytica Chimica Acta, 592:210-217

Sowokinos JR (2001) Biochemical and molecular control of cold-induced sweetening in potatoes. American Journal of Potato Research, 78:221-236

Stone JR (1974) Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society, 36:111-147

Strehmel N, Hummel J, Erban A, Strassburg K, Kopka J (2008) Retention index thresholds for compound matching in GC-MS metabolite profiling. Journal of Chromatography B-Analytical Technologies in the Biomedical and Life Sciences, 871:182-190

Steinfath M, Strehmel N, Peters R, Schauer N, Groth D, Hummel J, Steup M, Selbig J, Kopka J, Geigenberger P, Van Dongen JT. (2010) Discovering plant metabolic biomarkers for phenotype prediction using an untargeted approach. Plant Biotechnology Journal, 8:900-911

Stallard BR, Garcia MJ, Kaushik S (1996) Near-IR Reflectance Spectroscopy for the Determination of Motor Oil Contamination in Sandy Loam. Applied Spectroscopy, 50:334-338

Sulpice R, Trenkamp S, Steinfath M, Usadel B, Gibon Y, Witucka-Wall H, Pyl ET, Tschoep H, Steinhauser MC, Guenther M, Hoehne M, Rohwer JM, Altmann T, Fernie AR, Stitt M.(2010) Network analysis of enzyme activities and metabolite levels and their relationship to biomass in a large panel of Arabidopsis accessions. Plant Cell, 22:2872-2893

Syvanen AC (2005) Toward genome-wide SNP genotyping. Nature Genet 37: S5-S10

Szopa J, Wróbel M, Matysiak-Kata I, Swiedrych A (2001) The metabolic profile of the 14-3-3 repressed transgenic potato tubers. Plant Science, 161:1075-1082

Stushnoff C, Ducreux LJM, Hancock RD, Hedley PE, Holm DG, McDougal GJ, McNicol JW, Morris J, Morris WL, Sungurtas JA, Verrall SR, Zuber T, Taylor MA (2010) J. Exp. Bot., 61:1225-1238

Terzer M, Nathaniel D, Marcus WC, Jorg S (2009) Genome-scale metabolic networks, Wiley Interdisciplinary Reviews: Systems Biology and Medicine, 1:285–297.

Tibshirani R(1996) Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society, 58:267-288

Tieman DM, Zeigler M, Schmelz EA, Taylor MG, Bliss P, Kirst M, Klee HJ (2006) Identification of loci affecting flavour volatile emissions in tomato fruits. Journal of Experimental Botany, 57:887-896

Tikunov Y, Lommen A, de Vos CHR, Verhoeven HA, Bino RJ, Hall RD, Bovy AG (2005) A novel approach for nontargeted data analysis for metabolomics large-scale profiling of tomato fruit volatiles. Plant Physiology, 139:1125-1137

Tsuhako M, Nakajima A, Miyajima T, Ohashi S, Nariai H, Motooka I (1985) The reaction of cyclo-triphosphate with L-alpha alanine or beta-alanine. Bulletin of the Chemical Society of Japan, 58:3092-3098

Urbanczyk-Wochniak E, Baxter C, Kolbe A, Kopka J, Sweetlove LJ, Fernie AR (2005) Profiling of diurnal patterns of metabolite and transcript abundance in potato (Solanum tuberosum) leaves. Planta, 221:891-903

Urbany C, Colby T, Stich B, Schmidt L, Schmidt J, Gebhardt C (2012) Analysis of natural variation of the potato tuber proteome reveals novel candidate genes for tuber bruising. J Proteome Res, 11:703-16

Van den Berg RA , Huub CJ, Hoefsloot , Westerhuis JA, Age K, Smilde AK, Van der Werf MJ (2006) Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC Genomics, 7:142-157

Van Eck HJ and Jacobsen E (1996) Application of molecular markers in the genetic analysis of quantitative traits. I. In HJ Struik PC, Kouwenhoven JK, Mastenbroek LJ, Turkensteen LJ, Veerman a Vos J, ed, Abstracts of Conference Papers, Posters and Demonstrations of the 13th Triennial conference of the EAPR. European Association for Potato Research, Wageningen pp 130-131

Van Eck HJ, Jacobs JME, Van den Berg PMMM, Stiekema WJ, Jacobsen E (1993b) The inheritance of anthocyanin pigmentation in potato (Solanum tuberosum L.) and mapping of tuber skin colour loci using RFLPs. Heredity, 73:410-421

Van Eck HJ, Jacobs JME, Stam P, Ton J, Jacobsen E (1994) Multiple alleles for tuber shape in diploid potato detected by qualitative and quantitative genetic analysis using RFLPs.Genetics, 137:303-309

Van Ooijen J (2009) MapQTL(R) 6, Software for the mapping of quantitative trait loci in experimental populations

Vapnik VN (1995) The Nature of Statistical Learning Theory. Springer, New York.

Varma S and Simon R (2006) Bias in Error Estimation when using Cross-Validation for Model Selection. BMC Bioinformatics, 7, 91

Viksø-Nielsen A, Blennow A, Jørgensen K, Kristensen KH, Jensen A, Møller BL (2001) Structural, Physicochemical, and Pasting Properties of Starches from Potato Plants with Repressed r1-Gene. Biomacromolecules, 2:836-843

Viswanathan S, Unlu M , Minden JS (2006) Two-dimensional difference gel electrophoresis, Nature Protocols, 1:1351-1358

Visser RGF, Vreugdenhil D, Hendriks T, Jacobsen E (1994) Gene expression and carbohydrate content during stolon to tuber transition in potatoes (Solanum tuberosum). Physiol Plantarum, 90:285-292

Waaijenborg S and Zwinderman AH (2009) Sparse canonical correlation analysis for identifying, connecting and completing gene-expression networks. BMC Bioinformatics, 10:315

Waaijenborg S and Zwinderma AH (2007) Penalized canonical correlation analysis to quantify the association between gene expression and DNA markers, BMC Proceedings, 1:S122

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics, 10:57-63

Weckwerth W (2003) Metabolomics in systems biology. Annual Review Plant Biology, 54:669-89.

Weckwerth W, Wenzel K, Fiehn O (2004) Process for the integrated extraction identification, and quantification of metabolites, proteins and RNA to reveal their co-regulation in biochemical networks. Proteomics, 4:78-83

Wermuth N (1980) Linear recursive equations, covariance selection,and path analysis. J Amer Statist Assoc, 75:963-972

Werij JS, Kloosterman B, Celis-Gamboa C, de Vos CH, America T, Visser RG, Bachem CW (2007) Unravelling enzymatic discoloration in potato through a combined approach of candidate genes, QTL, and expression analysis. Theor Appl Genet, 115:245–252

Werij JS (2011) Genetic analysis of potato tuber quality traits, Ph.D. Thesis, Wageningen University, Wageningen, The Netherlands, 3pp, ISBN 978-94-6173-092-3

Wold H (1966) Estimation of principal components and related models by iterative least squares. New York: Academic Press. P.R. Krishnaiaah (Ed.) Multivariate Analysis, 391-420

Wold H (1975) Soft Modeling by Latent Variables; the Nonlinear Iterative Partial Least Squares Approach. Persp Prob Stat, Papers in Honour of M. S.Bartlett, ed. J. Gani, London: Academic Press

Wolters AMA, Uitdewilligen JGAML, Kloosterman BA, Hutten RCB, Visser RGF, Van Eck HJ (2010) Identification of alleles of carotenoid pathway genes important for zeaxanthin accumulation in potato tubers. Plant. Mol. Biol., 73:659–671

Wiesenborn DP, Orr PH, Casper HH, Tacke BK (1994) Potato Starch Paste Behavior as Related to Some Physical/Chemical Properties. Journal of Food Science, 59:644-648

Wienkoop S, Morgenthal K, Wolschin F, Scholz M, Selbig J, Weckwerth W. (2008) Integration of metabolomic and proteomic phenotypes: analysis of data covariance dissects starch and RFO metabolism from low and high temperature compensation response in Arabidopsis thaliana. Molecular Cell Proteomics, 7:1725-36

Xu X, Pan S, Cheng S, Zhang B, Mu D, Ni P, Zhang G, Yang S, Li R, Wang J, Orjeda G, Guzman F, Torres M, Lozano R, Ponce O, Martinez D, De la Cruz G, Chakrabarti SK, Patil VU, Skryabin KG, Kuznetsov BB, Ravin NV, Kolganova TV, Beletsky AV, Mardanov AV, Di Genova A, Bolser DM, Martin DM, Li G, Yang Y, Kuang H, Hu Q, Xiong X, Bishop GJ, Sagredo B, Mejía N, Zagorski W, Gromadka R, Gawor J, Szczesny P, Huang S, Zhang Z, Liang C, He J, Li Y, He Y, Xu J, Zhang Y, Xie B, Du Y, Qu D, Bonierbale M, Ghislain M, Herrera Mdel R, Giuliano G, Pietrella M, Perrotta G, Facella P, O'Brien K, Feingold SE, Barreiro LE, Massa GA, Diambra L, Whitty BR, Vaillancourt B, Lin H, Massa AN, Geoffroy M, Lundback S, DellaPenna D, Buell CR, Sharma SK, Marshall DF, Waugh R, Bryan GJ, Destefanis M, Nagy I, Milbourne D, Thomson SJ, Fiers M, Jacobs JM, Nielsen KL, Sønderkær M, Iovene M, Torres GA, Jiang J, Veilleux RE, Bachem CW, de Boer J, Borm T, Kloosterman B, Van Eck HJ, Datema E,Hekkert BL, Goverse A, Van Ham RC, Visser RG (2011) Genome sequence and analysis of the tuber crop potato. Nature, 475:189-95

Xu S and Hu Z (2010) Methods of plant breeding in the genome era. Genet Res, 92:423–441

Xu C, Ladouceur M, Dastani Z, Richards JB, Ciampi A, Greenwood CM (2012) Multiple regression methods show great potential for rare variant association tests.PLoS One, 8:e41694.

Yamada M, Greenham K, Prigge MJ, Jensen PJ, Estelle M (2009) The Transport InhibitorResponse 2(TIR2) gene is required for auxin synthesis and diverse aspects of plant development. Plant Physiology, 151:168:179

Yuan M and Lin Yi (2004) Model Selection and Estimation in Regression with Grouped Variables. Technical report, 1095

Yuan JS, Galbraith DW, Dai SY, Griffin P, Neal Stewart CJ (2008) Plant systems biology comes of age. Trends in Plant Science, 13:165-171

Yu TS, Kofler H, Hausler RE, Hille D, Flugge UI, Zeeman SC, Smith AM, Kossmann J, Lloyd J, Ritte G, Steup M, Lue WL, Chen JC, Weber A (2001) The Arabidopsis sex1 mutant is defective in the R1 protein, a general regulator of starch degradation in plants, and not in the chloroplast hexose transporter. Plant Cell, 13:1907-1918

Zamboni A, Carli MD, Guzzo F, Stocchero M, Zenoni S, Ferrarini A, Tonoi P, Toffali K, Desiderio A, Lilley KS, Pè ME, Benvenuto E, Delledonne M, Pezzotti M (2010) Identification of putative stage-specific grapevine berry biomarkers and omics data integration into networks. Plant Physiol., 154:1439-1459

Zeeman SC, Kossmann J, Smith AM (2010) Starch: its metabolism, evolution, and biotechnological modification in plants. Annual Review of Plant Biology, 61:209-234

Zeeman SC and Rees TA (1999) Changes in carbohydrate metabolism and assimilate export in starch-excess mutants of Arabidopsis. Plant, Cell & Environment, 22:1445-1453

Zhu H, Bilgi M, Snyder M (2003) Proteomics. Annual Review of Biochemistry, 72:783-812

Zou H and Hastie T (2005) Regularization and Variable Selection via the Elastic Net. Journal of the Royal Statistical Society: Series B, 67:301-320

Zou H (2006) The Adaptive Lasso And Its Oracle Properties. Journal of the American Statistical Association, 101:1418-1430

Summary

At Wageningen UR Plant Breeding large ~omics data sets have been collected over the years and different studies have been performed on phenotypic data sets from field trials in different years from the same genotypes of a segregating diploid potato population, known as the CxE population. This population has also been used extensively for mapping, QTL analysis and genotype*environment (GxE) interaction studies. The data which have been accumulated from this population includes molecular marker data, phenotypic data (*e.g.* developmental traits, tuber quality traits), microarray data, metabolomics data (LC-MS and GC-MS) and proteomics (2D DIGE) data.

The goal of this thesis was to investigate suitable statistical methodologies which can be used for analysis of ~omics data, for example, to relate a phenotypic trait to an ~omics data set and select a minimum set of markers (irrespective of being metabolite, transcript or protein) which together predict a quantitative trait. To investigate statistical methods, we compared them in terms of the accuracy of prediction of phenotypic traits, but also in terms of whether or not correlated variables are selected in groups to predict the phenotype, and in terms of being able to rank the variables with respect to their importance in the prediction of the trait, based on statistical properties. To achieve this goal we considered a number of different quality traits in potato (such as tuber flesh colour, enzymatic discoloration, phosphate content, cold sweetening traits, tuber shape and starch gelatinization) and ~omics data sets: transcriptomics, metabolomics (LC-MS & GC-MS) and proteomics data sets in the CxE potato population. The results of these studies are described in the five experimental chapters of this thesis.

In Chapter 2 we used and studied different regression methods to relate a quantitative phenotypic trait, tuber flesh colour of potato (as a response variable) to a metabolomics data set (as the predictor data set). We applied univariate regression and different regularized multivariate regression methods like ridge regression (RR), LASSO, elastic net (EN), principal components regression (PCR), partial least squares regression (PLS), sparse partial least squares regression (SPLS), support vector regression (SVR) and random forest regression (RF) to predict potato flesh colour from the metabolomics data set. We compared these methods in terms of mean square error of prediction, goodness of fit, variable selection and the ranking of the variables. In terms of the prediction error, elastic net performed better than other methods..

Further in this thesis, we propose a strategy to integrate transcriptomics and metabolomics (LC-MS) data sets and select a subset of the metabolites and transcripts which show an association with quality traits such as tuber flesh colour and enzymatic discoloration of potato (at different time points during storage; after peeling and exposure to the air). We used a Random Forest regression approach for this. Furthermore, a Pearson correlation

network reconstruction with traits (flesh colour, enzymatic discoloration), gene expression data and metabolites led to the integration of known as well as uncharacterized metabolites with genome regions associated with the carotenoid biosynthesis pathway. We show that this approach enables the construction of meaningful networks with regard to the known carotenoid pathway and also a putative connection with the flavonoid pathway which was not known before. We identified two novel putative non-volatile glucosides of carotenoid-derived volatile metabolites in the carotenoid pathway: 4,7-Megastigmadiene-3,9-diol-glucoside and 2,3-Dihydroxy-4-megastigmen-9-one-glucoside and they explain 45% and 9%, respectively, of tuber flesh colour variation.

From gas chromatography (Time-of-light) mass spectrometry (GC-TOF-MS) data sets we identified genetic factors underlying variation in primary metabolism in the CxE mapping population. We performed a QTL analysis for starch and cold sweetening related traits and we inferred links between these phenotypic traits and primary metabolites. We further applied RF regression to each of the starch related traits separately and metabolic profiles to determine the predictive power of metabolites for a given phenotypic trait. Using RF regression we found significant associations between phenotypic and metabolic traits. Putative predictors were tested and we confirmed their presence in an independent collection of potato cultivars. We found beta-alanine, an amino acid, significantly associated with the starch content in the CxE as well as in a non-related panel of different cultivars.

In order to obtain an insight in the relationships between proteins and flesh colour, enzymatic discoloration, starch and cold sweetening related quality traits, we used a proteomics data set and combined it with quality traits using genetic information through QTL co-localizations, as well as a correlation study between protein traits and quality traits. Results show pQTL hotspot areas in the genome, specifically chromosomes three, five, eight and nine. In this study a set of proteins was selected for further analysis and identification. The identification of proteins involved in the variation of important quality traits could generate new insights about the genetics of potato quality traits in relation to proteins and their expression. Only for a few of these proteins we were able to come up with a putative identity and in some cases this matched relatively well with a presumed function of the protein, for example in the case of enzymatic discoloration.

Finally, we performed am integrated analysis with gene expression, metabolite (LC-MS and GC-MS) and proteomics data from tubers of this diploid potato population and present an approach to reduce the number of co-expressed genes, metabolites and proteins. First, we used Random Forest regression to select subsets of the genes, metabolites (LC-MS and GC-MS) and proteins showing a significant association with different phenotypic traits including potato tuber flesh colour, enzymatic discoloration, tuber shape and starch gelatinization. Second, variation in expression of selected genes or concentration of metabolites and proteins are mapped, resulting in the identification of eQTLs, mQTLs and pQTLs across the genome. For each genomic region a single gene, metabolite and or protein is selected as representative for that region and used in an integrated network analysis. Such an integrated analysis not only results in a list of candidate genes and underlying metabolic pathways, possibly linking genes, metabolites and proteins, but also reveals interactions between different genomic regions underlying trait variation.

Samenvatting

Bij Wageningen UR Plant Breeding zijn er in de loop van de jaren grote datasets met ~omics-data verzameld en zijn er diverse studies uitgevoerd op fenotypische datasets verkregen uit veldproeven in verschillende jaren met dezelfde set genotypen van een splitsende diploïde aardappel-populatie bekend onder de naam CxE-populatie. Deze populatie is ook gebruikt voor genetische kartering, QTL-analyse en analyse van genotype*milieu-interactie (G*E). De gegevens die verzameld zijn betreffen moleculaire merker-gegevens, fenotypische gegevens (bijvoorbeeld ontwikkelingskenmerken en kwaliteitsgegevens van de knollen), microarray-gegevens, metaboliet-data (uit LC-MS en GC-MS) en proteomics-data (2D DIGE).

Het hoofddoel van dit proefschrift was om statistische methoden te onderzoeken op hun bruikbaarheid om ~omics-gegevens te analyseren, bijvoorbeeld om een fenotypisch kenmerk te relateren aan een ~omics dataset en een kleinste set van merkers (of het nu metabolieten, transcript-gegevens of eiwitgegevens betrof) te selecteren die gezamenlijk het kenmerk kunnen voorspellen. We hebben de methoden vergeleken voor wat betreft de nauwkeurigheid van de voorspelling van de kenmerken, maar ook hebben we bekeken of gecorreleerde variabelen al dan niet als groep geselecteerd worden, en of we vanuit de statistische eigenschappen de variabelen kunnen rangschikken naar hun belang voor de voorspelling van het kenmerk. Om dit doel te bereiken hebben we in de CxE-populatie een aantal kwaliteits-eigenschappen van aardappels bekeken (zoals vleeskleur van de knollen, enzymatische verkleuring, fosfaatgehalte, koude verzoeting, knolvorm en gelering van het zetmeel) en meerdere ~omics datasets: transcriptoom-gegevens, metabolieten (LC-MS en GC-MS) en proteomics. De resultaten van deze studies zijn beschreven in de vijf experimentele hoofdstukken van dit proefschrift.

In hoofdstuk 2 hebben we verschillende regressie-methoden gebruikt en vergeleken om een kwantitatief fenotypisch kenmerk, namelijk vleeskleur van aardappelknollen (als respons in de regressie) te relateren aan een metabolomics-dataset (de predictoren in de regressie). We hebben univariate regressie uitgevoerd en verschillende geregulariseerde regressiemethoden: ridge regressie (RR), LASSO, Elastic Net (EN), principale componenten-regressie (PCR), partial least squares regressie (PLS), sparse partial least squares regressie (SPLS), support vector regressie (SVR) en Random Forest regressie (RF). Deze zijn gebruikt om vleeskleur te voorspellen vanuit de metabolomics-dataset. We hebben de methoden vergeleken voor wat betreft hun fout in de voorspelling (mean square error of prediction), de goodness-of-fit, variabelen-selectie en de rangorde van de variabelen. In termen van de voorspelfout presteerde Elastic Net beter dan andere methoden. Vier variabelen die van belang waren voor de voorspelling van vleeskleur zijn onder voorbehoud geïdentificeerd als van carotenoïden afgeleide moleculen, afkomstig uit de carotenoiden-biosynthese-route.

Verder stellen we in dit proefschrift een strategie voor om trancriptomics en metabolomics (LC-MS) te integreren en een subset van metabolieten en transcripten te selecteren die een verband laten zien met kwaliteitskenmerken zoals vleeskleur van de knollen en enzymatische verkleuring (op verschillende tijdstippen tijdens de bewaring, na schillen en blootstelling aan de lucht). Hiervoor hebben we een Random Forest-aanpak gebruikt. Daarnaast leidde een reconstructie van een correlatie-netwerk met de Pearson correlatie van de variabelen met de eigenschappen tot een integratie van zowel bekende als onbekende metabolieten met genoomregio's die zijn geassocieerd met de carotenoïden-biosynthese-route. We laten zien dat deze aanpak het mogelijk maakt betekenisvolle netwerken voor wat betreft de bekende carotenoïden-biosynthese-route als ook een mogelijk niet eerder bekend verband met de flavonoïden-biosynthese-route. We hebben onder voorbehoud twee nieuwe niet-vluchtige glycosiden van carotenoïden afgeleide wel vluchtige metabolieten uit de carotenoïden-biosynthese-route geïdentificeerd: 4,7-Megastigmadiene-3,9-diol-glucoside en 2,3-Dihydroxy-4-megastigmen-9-one-glucoside, die respectievelijk 56% en 9% van de variatie in knolvleeskleur verklaren.

Vanuit gaschromatografie massaspectrometrie (GC-TOF-MS) datasets hebben we in de CxE karterings-populatie genetische factoren geïdentificeerd die ten grondslag liggen aan variatie in het primaire metabolisme. We hebben een QTL-analyse uitgevoerd voor zetmeel- en aan koude verzoeting gerelateerde eigenschappen en verbanden gelegd tussen deze kenmerken en de primaire metabolieten. Daarnaast hebben we Random Forest-regressie uitgevoerd voor elk van de zetmeel-gerelateerde eigenschappen afzonderlijk op de metaboliet-profielen om de voorspellende kracht van de metabolieten voor een gegeven eigenschap te bepalen. Met deze regressie vonden we significante verbanden tussen de fenotypische en metaboliet-gegevens. De mogelijk voorspellende variabelen zijn getoetst en we hebben hun aanwezigheid in een onafhankelijke collectie van aardappel-cultivars kunnen bevestigen. We vonden dat bèta-alanine, een aminozuur, een significante associatie had met zetmeel-hoeveelheid in de CxE-populatie zowel als in een niet-gerelateerde set van verschillende aardappelrassen.

Om een inzicht te krijgen in de verbanden tussen eiwitten en vleeskleur, enzymatische verkleuring, zetmeel- en koude verzoeting-gerelateerde eigenschappen, hebben we een proteomics-dataset gecombineerd met de kwaliteits-eigenschappen door gebruik te maken van genetische informatie (bij elkaar liggen van QTLs op het genoom) als ook een correlatie-studie tussen eiwit-profielen en kwaliteits-eigenschappen. De resultaten geven aan dat er hotspots voor eiwit-QTLs in het genoom zijn op chromosoom drie, vijf, acht en negen. In deze studie is een set voor kwaliteit interessante eiwitten geselecteerd voor verdere analyse en identificatie. De identificatie van eiwitten betrokken bij de variatie in kwaliteit zou nieuwe inzichten kunnen opleveren voor de overerving van kwaliteits-eigenschappen in relatie tot deze eiwitten. Tot nu toe konden we echter maar voor enkele eiwitten een voorlopige identificatie en dus mogelijke functie achterhalen. In sommige gevallen gaf de veronderstelde identiteit wel een aanknopingspunt naar een mogelijke bij de eigenschap betrokken functie van het eiwit, bijvoorbeeld voor enzymatische verkleuring.

Tenslotte hebben we een geïntegreerde analyse van genexpressie-, metaboliet (GC-MS en LC-MS) en proteomics-gegevens van de aardappelknollen van deze diploïde populatie uitgevoerd om te komen tot een kleinere set van gezamenlijk gereguleerde genen, metabolieten en eiwitten. Eerst voerden we een Random Forest-regressie uit om subsets genen, metabolieten en eiwitten te selecteren die een significant verband hadden met verschillende fenotypische kenmerken, waaronder vleeskleur van de knollen, enzymatische verkleuring, knolvorm en gelering van het zetmeel. Vervolgens hebben we de variatie in expressie van de geselecteerde genen, en in de concentraties van metabolieten en eiwitten gekarteerd op de genetische kaart van aardappel, hetgeen resulteerde in de identificatie van expressie-QTLs (eQTLs), metaboliet-QTLs (mQTLs) en eiwitQTLs (pQTLs) over het genoom. Voor elke genomische regio is een enkel gen, metaboliet en/of eiwit geselecteerd en gevisualiseerd in een geïntegreerde netwerk-analyse. Een dergelijke analyse levert niet alleen kandidaat-genen en mogelijk onderliggende metabolische routes op waarin genen, metabolieten en eiwitten gezamenlijk voorkomen, maar laat ook de mogelijke interacties zien tussen genoom-regio's betrokken bij de variatie in de eigenschappen.

Acknowledgements

At last I am "done" with my PhD thesis. It took me a longer time than expected but finally it is finished. In 2007, July 4th, I came to Wageningen to pursue my PhD after resigning from my job in India. So, obtaining a degree was very important for me. However, it would not have been possible if Prof. Richard G.F. Visser had not allowed me to do my PhD at Plant Breeding. I would like to thank Prof. Visser for giving me the opportunity to pursue my doctoral degree at the department of Plant Breeding under his supervision. This opportunity brought me not only an academic degree but also considerable personal development.

Dear Richard, I am very happy for your continuous support while I was away from Plant Breeding since July 2011. At times I was very worried about delays in finishing this Thesis, but you tried to keep up my spirit all the time. Thank you very much.

I would also like to thank my supervisor Dr. Chris Maliepaard. Chris, I still remember the day of July 4th, 2007 early morning when you came to Schiphol Airport to pick me up writing my name on a board. You did your best to make me feel at home by offering a helping hand all the time. You not only kept the door of your office open every time that I was coming with a question, but you also opened the door of your home for me and later on for my family.

Another person who helped me a lot to settling down in Wageningen is Joost van Langen. Joost, you tried your best to make me more familiar in The Netherlands and in Wageningen. Thanks for all the trips together with Chris and Paula.

I would also like to thank to my co-authors Natalia Carreno Quintero and Bjorn Kloosterman who helped me a lot to support my statistical results with proper biological confirmation.

I would not have done my Master's in India if I would not have got financial support from my friends Hrishikesh Dutta Gupta and Abanishwar Chakrabarti. Hrishi, I really appreciate the way you supported me in my Master's. I can never ever forget your contributions. Abanishwar, you always supported me all the time. Many thanks for everything!

Prashanta, Kaushik Saha (Kau), Rajat roy (Rajaji) and Shiladitya Da, you guys were all the time with me whenever I needed you. Many thanks for showing your strong friendship and bonding.

Two Indian families in The Netherlands who supported me in many ways are Pallabda and Boudi, Sagarika and Jayda, many thanks for your support while staying in The Netherlands.

Luis and Ricardo, thanks for everything, especially our visit to Paris in 2008. I hope we can get together some time.

Paula, I never forget our chit chat at the coffee machine on the second floor, thanks for helping me whenever I needed.

I would also like to thank some of my mentors, Prof. L.K. Arora, Prof. P.K. Dass, Prof. K.K Sasi and Prof. N. Yathindra who encouraged me during my Bachelor's and Master's days until today.

Then, a bunch of wonderful international friends were added to my list who made my PhD and my stay in Wageningen a wonderful time despite the difficulties: Shiva, Leila, Arwa, Nasim, Mirjana, Anna, Bjorn D‘Hoop, Brana, Nadeem, Sabaz, Ningwen, Nico, Pierre, Alireza, Yusuf, Xingfeng, Alvaro and Hulya.

My special thanks to Ram Kumar Basnet and Madhuri Madduri for being my paranymphs.

I cherished my moments with all my Indian friends Anitha, Madhuri and Manoj, many thanks for visiting us in Lyon, France and Gent, Belgium and boosting our spirit. Sameer and Vani, many thanks for your company especially when I broke my leg in Wageningen and you guys used to provide me food time after time. Geo and Sushimita, I cannot forget you people as a host in our Cologne visit for Durga Puja festival. Prasad, Abhishek, Rajshekar Korada, Geo, Sushimita, Kuppu, Parthada, Rama Bangera, Srinivasu (Mr.Kalaiah), Anand, and many more with whom I celebrated many dinners, birthday parties and graduation ceremonies.

My father is my true inspiration for going ahead in my career. He always said to me “only education can give a better life, especially to people in developing countries”. Baba and Maa, you have always supported, encouraged and believed in me for everything, even when I left my country and decided to go abroad to pursue my PhD.

I also would like to thank my father-in-law and mother-in-law who also supported me in various ways during this time.

Finally, last but not least, I would like to thank my wife Sharmistha Banik, who gave me our son Aniruddho and for the amount of trouble she shared in travelling to different countries due to changing jobs for all her help, inspiration and motivation, whenever I needed.

About the author

Animesh Acharjee was born on October 6, 1979 in Agartala, Tripura, North East India. After completing his schooling, he joined for a Bachelor degree in Technology in Electrical Engineering at the North Eastern Regional Institute of Science and Technology (NERIST), Itanagar, Arunachal Pradesh, India, where he completed in 2004. After finishing his Bachelor degree, he joined the Institute of Bioinformatics and Applied Biotechnology (IBAB), Bangalore, India for doing Master's studies in Bioinformatics. During this programme, he was awarded best student for "computational structural biologist". In July, 2007 he started his PhD program at Wageningen UR Plant Breeding, Wageningen University and Research Center (WUR), The Netherlands. This thesis presents the outcome of his PhD research on "**Systems biology and statistical data integration of ~omics data sets**". Presently, he is employed as a Systems Biologist at BASF-Cropdesign in Gent, Belgium.

Education Statement of the Graduate School

Experimental Plant Sciences

Issued to:	**Animesh Acharjee**
Date:	**12 June 2013**
Group:	**Plant Breeding, Wageningen University & Research Centre**

1) Start-up phase		*date*
►	**First presentation of your project**	
	~Omics data integration	Oct 12, 2007
►	**Writing or rewriting a project proposal**	
	PhD proposal: Systems Biology and Statitical data integration of ~omics data sets	Oct 08, 2007
►	**MSc/BSc courses**	
	ABG-30806 Modern Statistics for the Life Sciences	2008-2011
	MIB-31806 Systems Biology: From omics to integrative Biological Networks.	2008-2011
	PBR 20806 Plant Breeding	2008-2011
►	**Laboratory use of isotopes**	
	Subtotal Start-up Phase	*13.5 credits**
2) Scientific Exposure		*date*
►	**EPS PhD student days**	
	EPS PhD student day, Wageningen	Sep 13, 2007
	EPS PhD student day, University of Leiden	Feb 26, 2009
	EPS PhD student day, University of Utrecht	Jun 01, 2010
►	**EPS theme symposia**	
	EPS theme 4 'Genome Plasticity', University of Leiden	Dec 07, 2007
	EPS PhD Retreat (EPS, SVD, IMPRS), Wageningen	Oct 02-03, 2008
	EPS theme 4 'Genome Plasticity',Wageningen University, Wageningen	Dec 12, 2008
►	**NWO Lunteren days and other National Platforms**	
	GeneStat user day, 2008, Wageningen	Jun 18, 2008
	QTLMAS2009 workshop, Wageningen	Apr 20-21, 2009
►	**Seminars (series), workshops and symposia**	
	Symposium: Integrative bioinformatics : At the cuting edge of network analysis and biological data integration, Amsterdam	Nov 08, 2007
	Symposium: "Text mining for Dutch Genomics", Wageningen	Nov 23, 2007
	Genomic Researcher Event 2007, Amsterdam	Nov 29, 2007
	Research Day, laboratory of Plant Breeding , Wageningen	Jun 17, 2008
	Research Day, laboratory of Plant Breeding, Wageningen	Mar 03, 2009
	Research Day, laboratory of Plant Breeding, Wageningen	Feb 08, 2010
	Systems Biology Day, Wageningen	Jun 10, 2009
	MDA workshop with "penalty methods", Rotterdam	Oct 15, 2009
	Bio MDA meeting, Utrecht	Feb 18, 2010
	26th Symposium on Chemometrics, Utrecht	May 20, 2010
	Belgiam Chemometrics workshop on "Support Vector Regression", Belgium	Apr 30, 2010
	Systems Biology Day, Wageningen	Jun 16, 2010
►	**International symposia and congresses**	
	International workshop on "Probabilistic modelling in computational biology", Vienna, Austria	Jul 26, 2007
	RECOMB Regulatory genomics and systems biology, MIT, Boston, USA	Oct 29-Nov 02, 2008
	EUCARPIA, Biometrics in Plant Breeding Section, Scotland, UK	Sep 02-04, 2009
	Sixth Solanaceae Genome Workshop, New Delhi, India	Nov 08-13, 2009
	QTLMAS2010 workshop, Poznan, Poland	May 17-18, 2010
	Metabolomics2010, Amsterdam, The Netherlands	Jun 27-Jul 01, 2010
►	**Presentations**	
	Poster Presentation : EPS PhD student day, Wageningen Univerisity	Sep 13, 2007
	Poster Presentation : Research Day, laboratory of Plant Breeding	June 17, 2008
	Poster Presentation : Research Day, laboratory of Plant Breeding	Mar 03, 2009
	Poster Presentation : Research Day, laboratory of Plant Breeding	Feb 08, 2010
	Oral Presentation: EPS PhD Retreat, Wageningen	Oct 02, 2008
	Poster Presentation: MIT, Boston, USA	Oct 30, 2008
	Oral Presentation: Biometrics in Plant Breeding, Scotland, UK	Sep 04, 2009
	Poster Presentation: Sixth Solanaceae Genome Workshop, New Delhi, India	Nov 08-13, 2009
	Oral Presentation: QTLMAS2010 workshop, Poznan, Poland	May 17-18, 2010
	Oral Presentation: Metabolomics2010, Amsterdam, The Netherlands	Jun 30, 2010
►	**IAB interview**	Feb 17, 2011
►	**Excursions**	
	Subtotal Scientific Exposure	*23.3 credits**
3) In-Depth Studies		*date*
►	**EPS courses or other PhD courses**	
	Statistical learning and data mining ii: Tools for tall and wide data sets, Washington DC, USA	Oct 18-19, 2007
	PhD course: Systems Biology: ~omics data analysis (4 days)	Nov-Dec 2008
	PhD course: Multivariate Statistics (5 days)	May 11-13, 19-20, 2009
►	**Journal club**	
	Participation in a literature discussion group: Biometris (Systems Biology)	2007-2010
	Participation in a literature discussion group: Laboratory of Plant Breeding	2007-2011
	Participation in a literature discussion group: Department of Bioinformatics	2008-2011
►	**Individual research training**	
	Subtotal In-Depth Studies	*6.3 credits**
4) Personal development		*date*
►	**Skill training courses**	
	PhD assessment	Dec 04, 2007
	Techniques for writing and presenting a scientific paper Scientific writing	Apr 13-16, 2010
►	**Organisation of PhD students day, course or conference**	
	Organisation of meeting on "Regularaization"	2009-2010
►	**Membership of Board, Committee or PhD council**	
	EPS Board student member (2 years)	2008-2009
	Subtotal Personal Development	*4.4 credits**
	TOTAL NUMBER OF CREDIT POINTS*	**47,5**

Herewith the Graduate School declares that the PhD candidate has complied with the educational requirements set by the Educational Committee of EPS which comprises of a minimum total of 30 ECTS credits

** A credit represents a normative study load of 28 hours of study.*

Cover design:

By the Author

Printed by: CPI Wöhrmann Print Service, Zutphen, The Netherlands