Genome-wide transcriptomic profiling of turnip (*Brassica rapa* ssp. *rapa*) during tuber development

Niccolò Bassetti
890507037030

MSc Thesis Plant Breeding
(PBR-80436)

Supervisors:
dr.ir. Guusje Bonnema
dr.ir. Chris Maliepaard

July 2015
Laboratory of Plant Breeding
Wageningen University

Abstract

Turnip is a *Brassica rapa* morphotype that develops a tuber from the enlargement of the transition zone between hypocotyls and roots. Despite genetic and physiological studies little is known about the molecular factors determining turnip tuberization. The design of a custom *B. rapa* microarray allowed for the first time a genome-wide transcriptomic profiling during turnip tuber development. The highest variation in transcriptional changes occurred between 14 and 21 days after sowing (DAS), one week earlier than tuber initiation which occurred between 21 and 28 DAS. Weigthed Co-expression Network Analysis (WGCNA) was used to identify 16 modules of co-expressed genes characterized by distinctive profiles and that can be explored for future co-expression analysis. Modules with similar profiles were grouped and the resulting five clusters were used for pathway analysis using MapMan ontology. MapMan BINs showed significant enrichment for certain clusters and allowed to describe turnip tuber development. Turnip tuber accumulates sucrose, glucose, fructose but little starch. A detailed analysis of genes involved in the sucrose and starch metabolism was carried out to gain insights into the transcriptional regulation of the pathway.

Keywords: *Brassica rapa*, turnip tuber, transcriptomic profiling, pathway analysis, sucrose and starch metabolism

Table of Contents

Introduction

Brassica rapa

Brassica rapa L. (2n=2x=20) is one of the three diploid *Brassica* species and has been cultivated for many centuries all over the Europe and Asia (Dixon 2007). Natural and human selection for such geographically wide cultivation regions resulted in a rich abundance of morphotypes to be used as leafy vegetables, oilseed crops or edible tubers (Bonnema et al. 2011). Nonetheless, *B. rapa* genetic diversity has been associated more to the accessions geographical origin than to their morphological aspects, suggesting the involvement of few genes in the morphotype diversification (Zhao et al. 2005; Del Carpio et al. 2011).

Turnip (*Brassica rapa* ssp. *rapa*) is a morphotype characterized by an enlarged tuber and it is cultivated both as vegetable and fodder crop. The interest in turnip tuberization is placed in the context of the investigation on the genetics underlying the *B. rapa* morphological variation (Zhang et al. 2014). Moreover, research on turnip can also generate knowledge on the tuberization process itself and establish a comparison with the long-standing researches on other crops.

Research on turnip tuberization

Turnips develop tubers in the transition zone between root and hypocotyls although the contribution of the two tissues varies among different accessions (Takahashi et al. 1994; Zhang et al. 2014). The development and growth of storage organs involves a complex interaction of genetic, physiological and environmental factors.Genetic studies showed that the morphological variation for turnip tuber is under control of many quantitatively inherited traits with a complex genetic control. Quantitative trait loci (QTLs) for turnip tuber traits have been mapped across the genome on the majority of chromosomes (1, 2, 3, 4, 6, 7 and 9) (Lou et al. 2007; Lu et al. 2008; Kubo et al. 2010). However, none of these QTLs have been fine mapped yet.

Different *in vitro* studies have been carried out to determine the physiology of tuber development. The increase in radial dimensions of the tubers was primarily related to vascular cambium activity stimulated by sugars and phytohormones (Peterson 1973). In fact, the growth of excised tuber tip was promoted by a combination of auxins, cytokinins, sucrose and myo-inositol. Moreover, turnip epicotyls elongated through the continuous application of gibberellins (GA_3) were observed differentiating aerial tuber-like organs once the hormone application was suspended (Nishijima et al. 2005). Nonetheless, recent investigations showed opposite results and the role of hormones is still questioned. In fact, tuber initiation of turnip explants grown *in vitro* was not dependent upon addition of gibberellins (GA_3) and was inhibited by the addition of cytokinins (BAP) (Temesgen 2012; Zhang, personal communication). However, turnip explants could form tubers in presence of auxins and sucrose, confirming their promoting role. Further, turnip tuber development and filling have also been studied relatively to sucrose metabolism through enzymatic assays (Gupta et al. 2001). Despite these studies, a comprehensive understanding of the molecular and metabolic factors determining turnip tuberization is still lacking. Further research is required and additional knowledge can be derived from research on other crops.

Up to now, tuberization been more intensively investigated in crops like potato (*Solanum tuberosum*) (Xu et al. 1998; Jackson 1999; Abelenda et al. 2014; Kloosterman et al. 2013), radish (*Raphanus sativus*) (Ting and Wren 1980; Rouhier and Usuda 2001)), sweet potato (*Ipomea batatas*) (Wilson and Lowe 1973; You et al. 2003) and sugar beet (*Beta vulgaris*) (Lukaszewska et al. 2012). Research on those crops, especially in potato, have revealed some of the mechanisms involved in tuber formation and can provide additional knowledge for research in turnip. In potato, the transition from stolen to tuber is a photoperiod-dependent process sharing components with the flowering regulatory pathway (Abelenda et al. 2014). Particularly, a homologue of the flowering signal FLOWERING LOCUS T (FT), StP6A, acts as mobile tuberization signal under inducing conditions (Navarro 2011). The role of hormones is also fundamental. Tuberisation is induced from reduced level of gibberellins in the sub apical region of the stolen (Vreugdenhil and Sergeeva 1999). Instead, auxins and strigolactones have an antagonistic effect on onset of potato tuber, with auxin promoting tuber formation and strigolactones repressing bud outgrowth (Roumeliotis et al. 2012).

Genomic approaches

The recent sequencing of the *B. rapa* genome offered the opportunity to broaden the research on the morphological diversification of *B. rapa* species and so on turnip (Wang et al. 2011). The evidence of a whole genome triplication after the divergence from *Arabidopsis thaliana* suggested that multiple orthologs of the same gene may have undergone neo- or sub-functionalization facilitating the emergence of genetic factors responsible for the extreme morphological variation (Cheng et al. 2014). Further, genomic resources allowed comparative genomic studies to investigate selection signatures across different morphotypes. Recently, a DH line of turnip accession VT_117 was re-sequenced, assembled and annotated to be compared with the *B. rapa* reference genome from the Chinese cabbage Chiifu (Lin et al. 2014).

The availability of the genome sequence allows the investigation of transcriptome, that means of the gene expression at the genome level. While transcriptomic profiling with microarray have been carried out to study seed development of *B. rapa* (Basnet et al. 2013), this work represents the first attempt on developing turnip tuber. Transcriptomic profiling has been widely applied to many crops to obtain valuable insights into developing plant tissue or organs. Relatively to storage organ, transcriptomic studies have been carried out on potato tubers (Kloosterman et al. 2005; Kloosterman et al. 2008), sweet potato tuberous root (Firon et al. 2013) and radish (Mitsui et al. 2015).

Moreover, transcriptomic profiling allowed the so called genetical genomics studies (Jansen and Nap 2001). In fact, transcript abundance is a heritable trait and can be measured in a segregating population to map *expression quantitative trait loci (eQTL)* that can help to disentangle the transcriptional regulatory mechanisms of complex trait. In this sense, transcriptomic profiling of a developing organ may provide valuable information to be used for further expression studies or genetical genomics approaches.

Objectives

The aim of the present work is to characterise the turnip tuber formation through analysis of transcriptomic profiling data from microarray. The four main objectives of the analysis are:

1) obtaining a global overview of the transcript abundance across the tuber developmental stages;
2) defining modules of co-expressed genes and characterize their expression profiles;
3) investigating co-expression modules for over- (or under-) presence of gene functional categories;
4) investigating the role genes of functional categories known to be involved in turnip tuber development (carbohydrates, hormones, etc.).

Materials and Methods

The present work involved only the analysis of microarray data. The experimental procedures that generated this data were carried out by dr. Ningwen Zhang. Data analysis was carried out on the statistical software environment R (Ihaka and Gentleman 1996).

Plant Material

A doubled haploid (DH) line DH-VT_117 was chosen from DH lines generated from the Japanese vegetable turnip accession VT_117 (CGN15201). The genome of DH-VT_117 has been recently resequenced, re-annotated and compared with the *B. rapa* reference genome form Chinese cabbage Chiifu and the resequenced oil type RC_144 (Lin et al. 2014). The turnip VT_117 is characterized by round-shaped, red peel tuber, primarily composed by enlarged hypocotyls, and early flowering (60-80 days after sowing). The plants were grown with a standard pot soil in a climate chamber, with 20/18 °C day/night temperature and 16h/8h day/night length. The experimental design consisted of two biological repeats grown over 6 time points: 7, 14, 21, 28, 35 and 42 days after sowing. Each biological repeat was obtained by pooling three different turnip tubers. Hypocotyls tissues were removed from the plants and immediately immersed into liquid nitrogen to prevent RNA degradation.

RNA isolation

RNA isolation was done using RNeasy mini kit according to the manufacturer's instructions (Qiagen, Milden, Germany) followed by DNase treatment (AmpGrade I, Invitrogen, Burlington, ON, Canada) and a purification step (RNeasy Mini Kit, Qiagen). The quantity of RNA was determined by NanoDrop ND-100 UV–VIS spectrophotometer and quality was assessed by A260/A280 and A260/A230 ratio (NanoDrop Technologies, Inc., Wilmington, DE, USA) as well as by 1% agarose gel.

Microarray probe design and hybridization

A custom microarray was designed using the whole genome sequence of *B. rapa* cv. Chiifu (morphotype Chinese cabbage) version 1.0 (Wang et al. 2011). The predicted genes models of the genome sequence were used to design oligonucleotides probes (60-mer) for a two-colour Agilent microarray platform. The microarray contains 61,551 probes which represent 39,498 *B. rapa* genes assigned to one of the ten chromosomes and 1406 *B. rapa* genes assigned to scaffolds. Of the total 40,904 *B. rapa* genes, 20,647 were represented on the microarray by two probes. These two probes are different as the starting nucleotide of the 60-mer differs. Nonetheless, in the following text they are defined as "duplicated" probes as designed on the same gene.

Cy3 and Cy5 dyes were incorporated into cDNA samples according to the Agilent two-colour microarray based gene expression analysis (Low input quick Amp labelling G4140-90050) protocol (Agilent Technologies, Inc., Santa Clara, CA, USA) and hybridized on arrays following a self-self design. To obtain this design, technical replicated of each sample were labelled with Cy3 and Cy5and hybridized on the same array. In total twelve hybridizations were carried out consisting of samples from six time points, each having two biological replicates. Slides were scanned and raw expression values were extracted using Agilent Feature Extraction software.

Pre-processing

Raw expression values of the microarray were acquired with Agilent Feature Extraction software. No background correction was applied as no major benefits were previously found for Agilent platforms (Zahurak et al. 2007). The signals of the two dyes, red (R) and green (G), were transformed by a logarithm base 2 and the spot for each gene *i* was represented in terms of expression ratios M and expression intensity A:

$$\mathrm{M}_i = \log_2 \frac{R}{G} = \log_2 R - \log_2 G \qquad (1)$$

$$\mathrm{A}_i = \frac{(\log 2\mathrm{R} + \log 2\mathrm{G})}{2} \qquad (2)$$

The advantage of logarithmic transformation is that it treats numbers and reciprocals symmetrically. For the same gene *i*, an up-regulation of 2-fold results in result in $\mathrm{M}_\mathrm{i} = \log_2(2) = 1$ while a down-regulation of the same magnitude will be $\mathrm{M}_\mathrm{i} = \log_2(0.5) = -1$, thus giving equal weight to different direction of gene regulation. Normalization within slides was carried out with locally weighted linear regression (lowess) to correct for many source of variation, namely the different labelling efficiencies of the two dyes. Lowess is a common scatter plot smoother (Cleveland 1979). A non-parametric regression is fitted for each data point to emphasize the effect of the neighbour data points. The fitted values are then subtracted from the M-values of each spot according to the formula (Smyth and Speed 2003):

$$\mathrm{N}_i = \mathrm{M}_i - \mathrm{loess}(\mathrm{A}_i) \qquad (3)$$

While normalization within slides shifts M-values around zero by correcting for intensity and spatial bias, normalization between slides accounts for variation due to differences between arrays. Normalization between slides was carried out by means of quantile normalization (Bolstad et al. 2003). The method enforces the distribution of probe intensities of each array to be the same across arrays. Give a dataset of *n* arrays (columns) with *p* probes (rows): 1) each column is sorted; 2) for each row is calculated the mean; 3) the mean replace the values of each element of the row, 4) the correct column are rearranged to have the same order as the initial matrix. Due to the self-self design, both signals R and G measured the transcript abundance detected by each probe. Therefore, the normalized A values were used for downstream analysis. All the pre-processing steps were carried out with R-Bioconductor package *limma* (Smyth 2005).

Principal components analysis

Principal components analysis (PCA) is a data reduction procedure commonly applied to large multivariate datasets. PCA was applied log transformed and normalized expression values of the all dataset. The procedure was implemented with the *pricomp* function of R using the covariance matrix without scaling the data.

Weighted gene co-expression analysis

Weighted Gene Co-expression Network Analysis (WGCNA) is a method that allows to construct a correlation-based network of modules composed by co-expressed genes (Zhang and Horvath 2005). The network approximates a scale-free topology where the distribution of the nodes connectivity *k* follows the power law $p(k)\sim k^{-\gamma}$. This determines that few nodes (genes) are highly connected and act as hubs of co-expressed modules. First, a similarity matrix s_{ij} is computed using pairwise correlations between genes using Pearson correlation or methods robust to outliers such as Spearman correlation or biweight midcorrelation (Wilcox 2012; Langfelder and Horvath 2012). The similarity matrix is defined "unsigned" (3) or "signed" (4) based on whether the correlations are calculated in absolute terms or not. s_{ij} is then transformed in adjacency matrix a_{ij} through a function that highlights strong correlations and penalizes the weaker ones on an exponential scale (5).

$$s_{ij}^{\mathrm{unsigned}} = |cor(x_i, x_j)| \qquad (3) \qquad \text{or} \qquad s_{ij}^{\mathrm{signed}} = \frac{1+|cor\,(\mathrm{xi,xj})|}{2} \qquad (4)$$

$$a_{ij} = |s_{ij}|^{\beta} \qquad (5)$$

The choice of the β parameter is based on the scale free topology criterion (Zhang and Horvath 2005).The adjacency information is further transformed into a topological overlap matrix (TOM) (6).

$$\mathrm{TOM_{ij}} = \frac{\sum_u a_{iu} a_{uj} + a_{ij}}{\min\{\sum_u a_{ju}, \sum_u a_{ju}\} + 1 - a_{ij}} \qquad (6)$$

$$\mathrm{DissTOM_{ij}} = 1 - \mathrm{TOM_{ij}} \qquad (7)$$

TOM uses the adjacency information to highlight connections shared between pair of genes *i* and *j* with all other genes *u* of the matrix. This method was found to give co-expressed modules with preserved biological information (Ravasz et al. 2002). It can be shown that $0 \leq \mathrm{TOM}_{ij} \leq 1$, then TOM can be subtracted to 1 to obtain a dissimilarity measure that range between 0 and 1(6) (Zhang and Horvath 2005). DissTOM was used as distance measure for hierarchical clustering. Modules were defined with the dynamic tree cut algorithm included in WGCNA package (Langfelder et al. 2008). The algorithm cuts the dendrogram at variable heights and allows a higher number of objects to be assigned to a cluster compared to constant height cut-off values. Each module can be summarized with the "Module Eigengene" (ME), the first principal component calculated using all the probes in the module (Horvath and Dong 2008). Paerson correlation between MEs can be used to merge similar module preserving the biological information. WGCNA method was implemented using the relative R package (Langfelder and Horvath 2008).

Functional annotation and pathway analysis

Pathway analysis (also enrichment analysis) is a common analysis that allows to define the biological meaning of the extensive lists of genes resulting from genomic data without the burden of a tedious manual search (Tipney and Hunter 2010). Although different methods have been elaborated, a common practice is to assign a functional annotation to the genes and test the annotated terms for under- or over-representation in a subset of data with the Fisher's exact test. Functional annotation of the probes was done using the MapMan ontology, a plant specific gene ontology (Thimm et al. 2004). MapMap is organized in 35 functional categories (BINs) and the file with the mappings between *B. rapa* genes and MapMan categories was downloaded at http://mapman.gabipd.org/web/guest/mapmanstore. Enrichment of MapMan BINs in WGCNA modules was calculated using the probes retained for WGCNA analysis as background and the probes of the module as subset. An example of contingency table is given below. Fisher's exact test calculates a probability (p-value) for the frequency of the annotated term in the subset with the frequency expected by chance given his occurrence in the background. A cut-off (eg p-value <0.05) is set to define term as enriched.

Differential expression

Differential expression analysis was used to select genes with expression values significantly different across time points. From the complete dataset, negative controls and 66 probes that could not be mapped to the *B. rapa* genome v1.0 were removed. Further, probes with expression values lower than the 95th percentile of negative control expression values were considered not reliable and were also filtered out. For each the resulting 51,469 probes a linear model was fit with the package *limma*. Difference in expression values between consecutive time points were tested with the empirical Bayes t-test (Smyth 2005). Significant probes between time points were defined at false discovery rate (FDR) <0.05 and log fold change >1.5 (Benjamini and Hochberg 1995).

Results and Discussion

Pre-processing of the data

The dataset presented expression values for all the probes with no missing values or bad quality spots. The quality of the microarrays data was assessed before normalization through the inspection of box-plots and smoothed densities plots of M- and A-values of raw expression values for each array (*Figure 1*). M-values presented bimodal distributions around positive modes as expected due to the self-hybridization design and the known dominant fluorescence of the red over the green dye. Nonetheless, M-values showed also negative values due to higher G signal compared to R. This was the result of the known gene specific dye bias that determines a different incorporation of the two dyes for different genes (Dobbin et al. 2005). However, the self-self design determined the use of the A-values as a measure of transcript abundance, therefore the probe-dye interaction bias was not further investigated. The distributions of the A-values were similar across arrays and presented skewness to the left, with the majority of the probes expressed at low intensities.

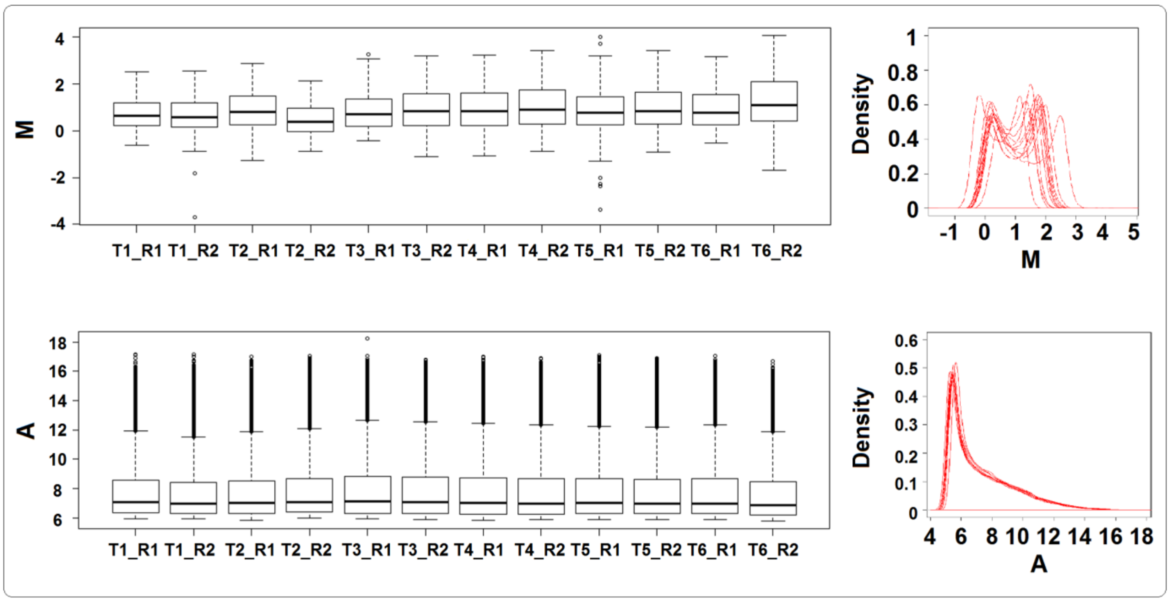


Figure 1. Box-plot and smoothed densities plot of M (upper) and A (lower) values for each of the twelve arrays. Each array is labelled as “time point T of replicate R ”.

The dependency of the log ratios M respect to the log intensities A was assessed through a MA plot, a scatter plot commonly used as diagnostics to visualize microarray signals (*Figure 2*). The general trend for all the arrays was a non-linear increase in expression values of the red over the green signal with the increase of the intensity. Spike-in and control probes are highlighted by different colors. The position of Agilent’s negative controls (SLV1, yellow), and non-organism spike-in (ERCC, dark green) on the lowest intensities suggests a good quality of the hybridization.

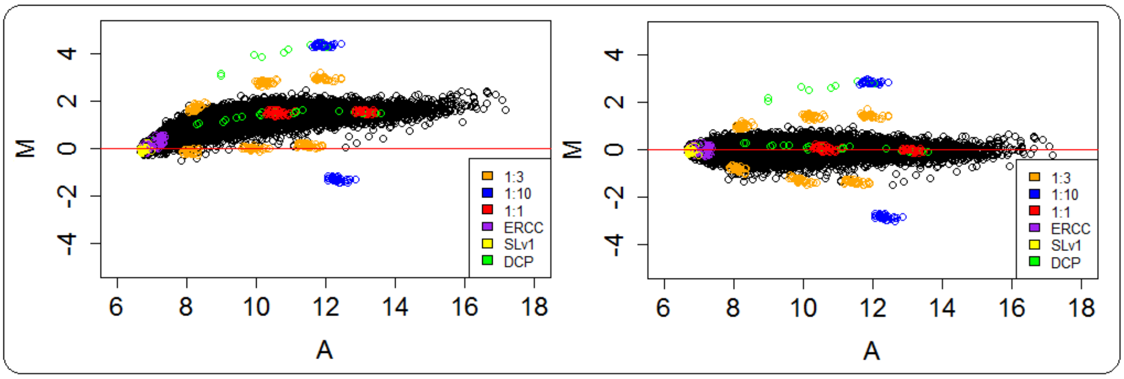


Figure 2. MA plot of the first array, before (left) and after lowess normalization within array (right).

Lowess regression was applied separately to each array to correct the trend of the M-values. Although the self-hybridization design determined the use of the intensity A as expression for downstream analysis, normalization within array was still recommended to reduce the noise during the execution of the between array normalization. Lowess had effect on log ratios and the results was a centring of the MA plot around M=0 (*Fig. 2*). Further confirmation of the good hybridization results from the position of the control probes with a titration series (1:1, red; 1:3, orange and 1:10, blue). The ratio represented different quantity of RNA labelled with the two different dyes and the expected position. In fact, a ratio of 1:1, 1:3 and 1:10 were respectively found at $\log_2(1)= 0$, $\log_2(3) = 1.58$ and $\log_2(10) = 3.32$. Similarly, the reciprocals of the titration ratios were found at negative values of the log values of the ratios. Furthermore, the application of the quantile normalization enforced the A signal of each array to assume the same distribution allowing a better comparability of the arrays in absence of connected design.

The normalized A signal was then used to identify samples outliers and to assess the quality of the biological replicates. Multivariate dataset characterized by the "small *n* large *p*" are particular sensitive to sample outliers which may interfere with the real biological signal within the experiment. A hierarchical clustering with Euclidean distance was applied with the UPGMA algorithm on the twelve arrays (*Figure3*).

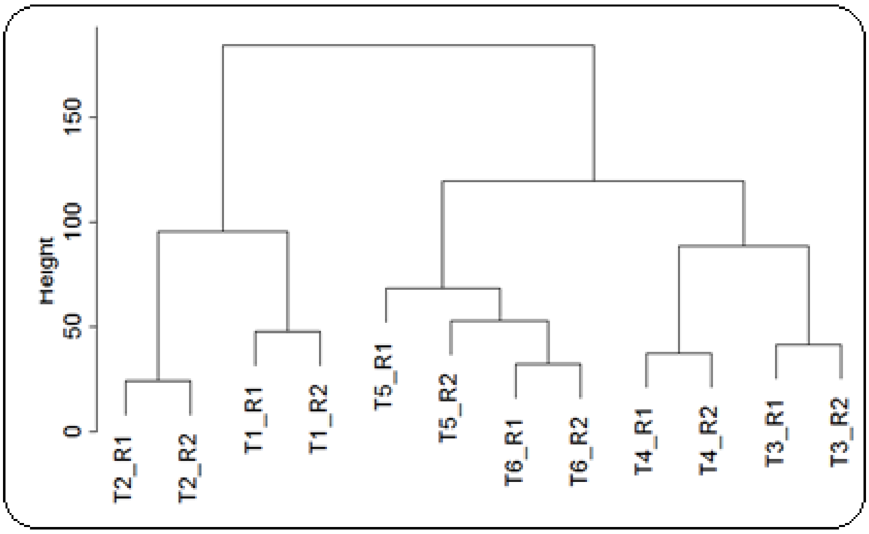


Figure 3. Hierarchical clustering of the twelve samples. Each array is labelled as "time point T of replicate R ".Biological repeats cluster together at each time point showing no sample outlier.

No outliers were found as the biological replicates were clustered together at each time point. Clustering results were confirmed by the high correlation of biological replicates across time points, showing Pearson correlation coefficients greater than 0,99 (Appendix A, *Table 1*).

The microarray presented 20,257 couple of duplicated probes (see Methods for a definition of duplicated probes) out the total 61,551 mapping to annotated *B. rapa* genes. Pearson correlation coefficients (PCC) between duplicated probes were calculated to investigate the validity of the custom array. The distribution of all PCC is shown in *Figure 4*. The two biggest bins contained 46.9% of the total duplicated probes (9,679/20,647) and showed PCC higher than 0.8 across time points, suggesting that they were actually mapping the same gene. However, the remaining 53.9% (10,968/20,647) of the duplicated probes showed decreasing values of PCC between 0.8 and -1. Also, 14.5% (3,011/20,647) had negatively correlated expression profiles across the six time points. Assuming the good design of the probes, this result might be explained through alternative splicing (AS) of *B. rapa* genes. However, a recently comprehensive analysis of *B. rapa* transcriptome with RNAseq found AS events only on 7,688 genes (Tong 2013). Therefore the phenomenon of AS cannot entirely explain the behaviour of the duplicated probes on the custom microarray. Alternatively, it can be argue the wrong design of some microarrays probes since based on predicted gene models. Overall, in absence of a clear reason to motivate the exclusion of low correlated probes, all the probes were retained for further analysis, being aware that this may have resulted in probes mapping the same genes but clustered in different co-expression modules.

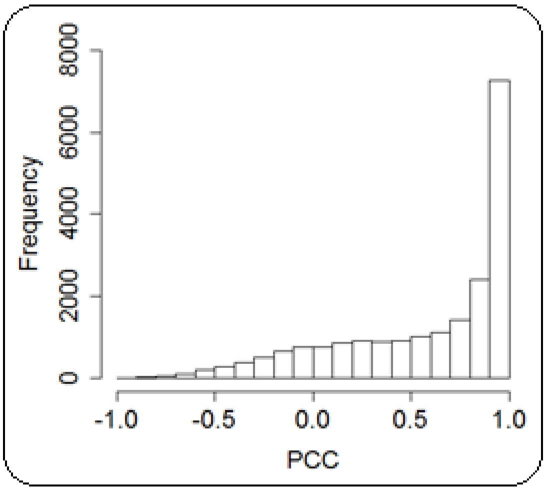


Figure 4. Distribution of Pearson correlation coefficients (PCC) among duplicated probes on the array. Probes are grouped in bins of 0.1. The y-axis displays the number of probes while the x-axis displays the PCC from -1 to 1.

As microarrays determine transcript abundance through hybridization, the measurements of gene expression lack of sensitivity and accuracy compared to traditional non high-throughput molecular biology techniques. Thus, a recommended practice is to validate microarray results through qRT-PCR in order to establish correlations between the two measurements. The RNA samples used in this project were also used for qRT-PCR experiment on 21 candidate genes (Habtemariam 2012
). As that dataset was not available,7 out 21 genes were chosen for a visual inspection and comparison of the expression profiles from microarray with a similar plot showed in Habtemariam (2012). The two plots are shown in *Figure 5*. Overall, the expression of the seven genes as detected with the microarray resulted similar to that one detected through qRT-PCR. A difference in the magnitude of the fold change can be ascribed to the difference in accuracy of the two methods.

Altogether, these results suggested a successful hybridization and a general good quality of the arrays. The biological replicates were consistent across time points and sample outliers were not found. A minor presence of low or even negative correlations between duplicated probes could be caused by alternative splicing on *B. rapa* gene, wrong design of the oligonucleotides or inaccurate prediction of gene models in the genome. Since it was not possible to discriminate among these possible causes, all the duplicated probes were kept for further analysis.

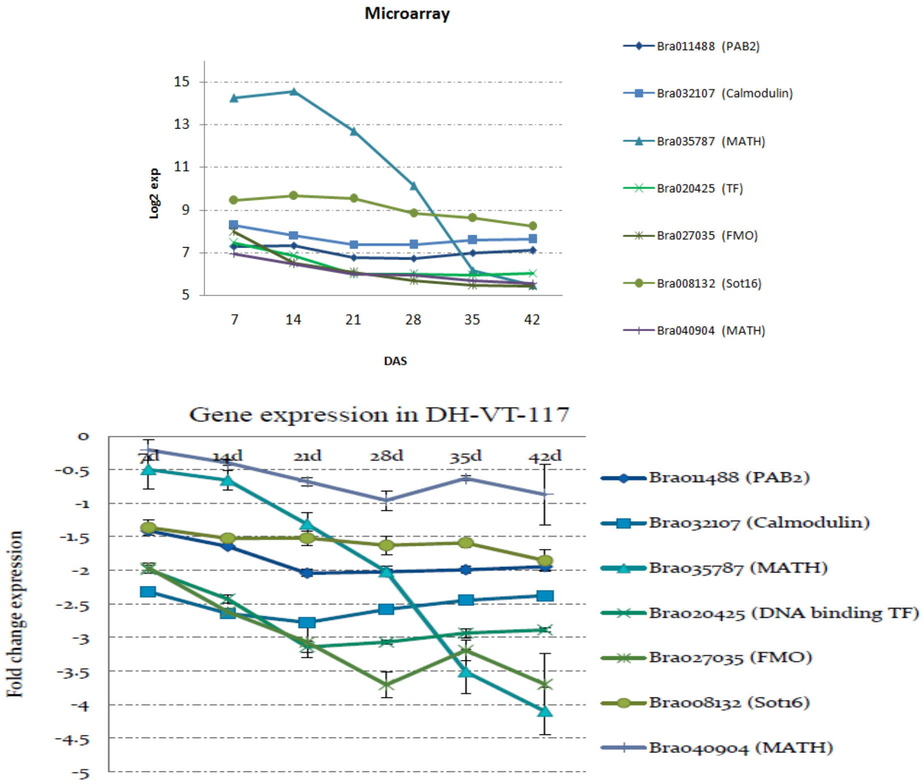


Figure 5. Indirect assessment of the reproducibility of the microarray through comparison of the expression profiles of seven genes as measured with the microarray (upper) and with RT-PCR (lower) by Habteramiam (2012).

Global overview of transcript abundance

All the 62,800 microarray probes showed binding to transcripts across all samples with no missing values. Control probes and probes that could not be mapped on the *B. rapa* genome v1.0 were excluded, resulting in a final dataset of 61,461 probes which detected transcripts of 40,838 *B. rapa* genes. 37,713 genes present homology with *Arabidopsis thaliana* genes, while the remaining 3,125 are considered unique gene of *B. rapa* species.

Turnip plants harvested at each time point of the experiment are shown in *Figure 6A.*Turnip growth during 42 days presented two different phases as major morphological modifications appeared only at 28 days after sowing (DAS). A first phase corresponded to the three earlier time points (7, 14, 21 DAS) and was determined by seed germination, seedling establishment and elongation. The second phase started with an established tuber (28 DAS) that underwent changes in length and thickness in the later time points (35 and 42 DAS). Thus, turnip tuber initiation is likely to take place between 21 and 28 DAS.

The profiles of the retained probes give a global overview of the transcriptome during the six time points (*Figure 6B*). Overall, two phases can also be distinguished regarding the dynamics of transcript abundance. The wider range of intensities detected at earlier stages (7 DAS) declines progressively till 21 and 28 DAS (time point 3 and 4). This phase corresponds to the morphological changes from the germination of the seed until the enlargement of the tuber (28 DAS). Then, the range of transcript assumes an increasing range of intensities until 42 DAS. This second phase corresponds to the growth of the tuber through filling with reserve substances (*Figure 6A*). The global overview of all probes profiles suggested that turnip transcriptome encounters massive changes in expression profiles at the interval 21-28 DAS, corresponding to major morphological changes in turnip tuber. Similarly, global coordinated change of gene expression profiles in correspondence of major developmental stages has been observed in other systems as *Arabidopsis* (Schmid et al. 2005), grape (Deluc et al. 2007) and wheat (Wan et al. 2008).

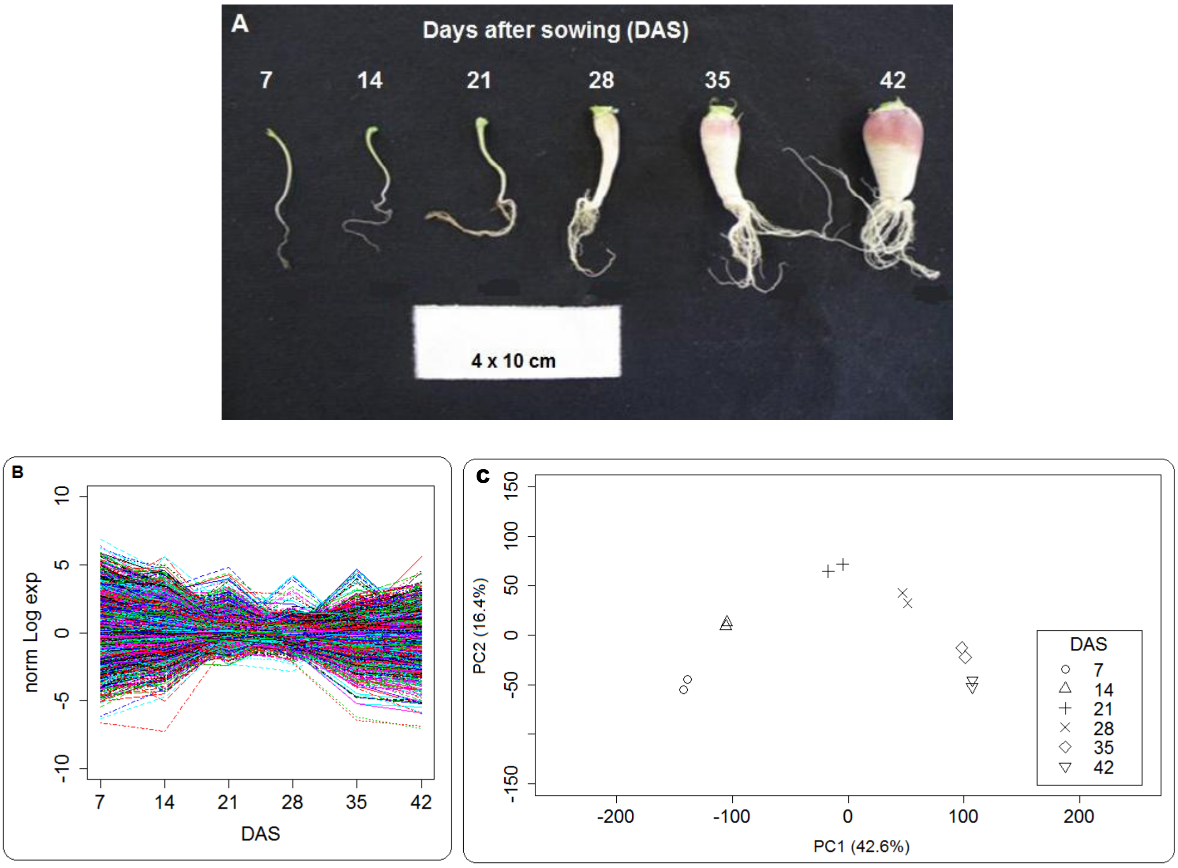


Figure 6. A) Development of turnip tuber at 7, 14, 21, 28, 35 and 42 days after sowing (DAS). Two plants were harvested at each time point and hybridized to the custom microarray. B) Global overview of the profile of all transcripts of turnip. Expression values are normalized to the mean to facilitate the visualization. C) Principal components analysis (PCA) based on transcriptional profiles during turnip development (7 – 42 DAS).

In order to assess the quality of the biological replicates and to obtain an overview of the global variation in transcript abundance, Principal components analysis (PCA) of the 61,461probes was performed (*Figure 6C*). The first two principal components (PC1 and PC2) captured 59% of the total variance, with 42,6% and 16,4% respectively (Appendix B *Table 1*). In accordance with the results of the clustering of the samples, biological replicates were found grouping together. PC1 discriminates the different time points, with the highest variation at the earlier time points (14-21 DAS). Basnet et al. (2013), while studying the transcriptome profiling of seed development in *B. rapa* pack choi and oil type, suggested to indicate the two time points with the highest distance on the PCA biplot as the right timing to carry out genetical genomics studies. However before drawing similar conclusion in turnip, it should be investigated if the variation in transcript abundance observed in the present work has similar timing in other turnip accessions.

Identification of co-expression modules

Genes transcribed with similar expression profiles often belong to similar or connected pathways. Clustering co-expressed genes is common practice to infer associations between profiles and biological functions allowing to reconstruct such pathways. Clustering was implemented using the WGCNA algorithm, a method which performs hierarchical clustering on a dissimilarity matrix calculated by Topological Overlap (TO) distance. Since TO is based on the computation of pairwise correlation between probes, a subset with the 20% most variable probes (10,280 out of 61,461 , mapping 7,654 genes) was retained to reduce the computational burden. The choice of the β parameter, which is required to construct the adjacency matrix, was carried out upon analysis of the fit of the network to the scale-free topology criterion (see Methods). The attempts to construct an unsigned network failed for reasonable values of β. A signed network was then built after defining a value of 18 for β (Appendix B, *Figure 1*). Probes were clustered together and a total of 56 co-expression modules were identified using a dynamic tree-cutting algorithm in order to assign all the probes to a cluster (Appendix B, *Figure2*). A minimum cluster size was set to 50 so that one of the 56 modules resulted composed by probes that could not be assigned to the any other modules. The global profile of the modules was summarized by "Module Eigengene (ME)" and MEs correlated above 0.9 were grouped together to obtain a final number of 16 modules (Appendix B, *Figure 2*). Each module was characterized by a free-scale topology with few hub genes highly connected to the other genes. Hub genes define biological significant modules and are often key regulatory genes of pathways (Horvath and Dong 2008). However, as it was decided not to pursue with the co-expression analysis, modules were organized in five major clusters and assigned to each of the cluster if exhibiting clear similarities in expression patterns.

The 16 co-expression modules identified with the WGCNA method are presented in *Figure 7*. The five clusters are numbered by Latin number I-V and displayed on the right side of the image. Modules showed difference in patterns and size, ranging from a minimum of 56 to a maximum of 4,710 probes. 492 probes (4.78% of 10,280) could not be assigned to any cluster and were included in module 1 ("grey", cluster I). The majority of the probes (4,710, 45.81%) were assigned to module 2 and were expressed in a steady-state down-regulated manner. Instead, three modules (3 to 5) showed a decline in transcript abundance not constant. In these modules the expression was rapidly down-regulated until 21 DAS and then appeared more stable. Modules 2-5 were grouped in cluster II. A total of 692 probes (6.73%) were included in four modules (6 to 9, cluster III) characterized by transient peak increases in the expression profiles. The peaks of expression showed a gradually delayed timing between module 6 and 8 (14-21 DAS), module 7 (21 DAS) and module 9 (21-28 DAS). Conversely, three modules presented a transient down regulated profile (10 to 12, cluster IV). Module 10 (120 probes, 1.16%) was down regulated between 14 and 28 DAS while module 11 (56, 0.54%) peaks at 21 DAS and module 12 (156, 1.51%) at 21-28 DAS. A group of modules were identified by probes with an upregulated profile (13 to 16, cluster V). Module 14 was the second biggest module (1930 probes, 18.77%) and presented a continuous upregulation from 7 to 42 DAS. Instead, module 15 (689 probes, 6.7%) was characterized by stable expression at earlier stages (7-14 DAS), up-regulation (14-35 DAS) and stable expression at later stages (35-42 DAS). Module 13 (56, 0.54%) also showed up-regulated expression although with a transient peak at 14 DAS. Finally, module 16 (548 probes, 5.33%) showed upregulation till 28 DAS and stable expression at later stages (35-42 DAS). Given these results, cluster analysis revealed that transcriptional expression during turnip tuber formation and development is not only a progressive process, where genes show a continuous up- or down-regulation, but also a dynamic process with genes showing transient variation in transcript abundance.

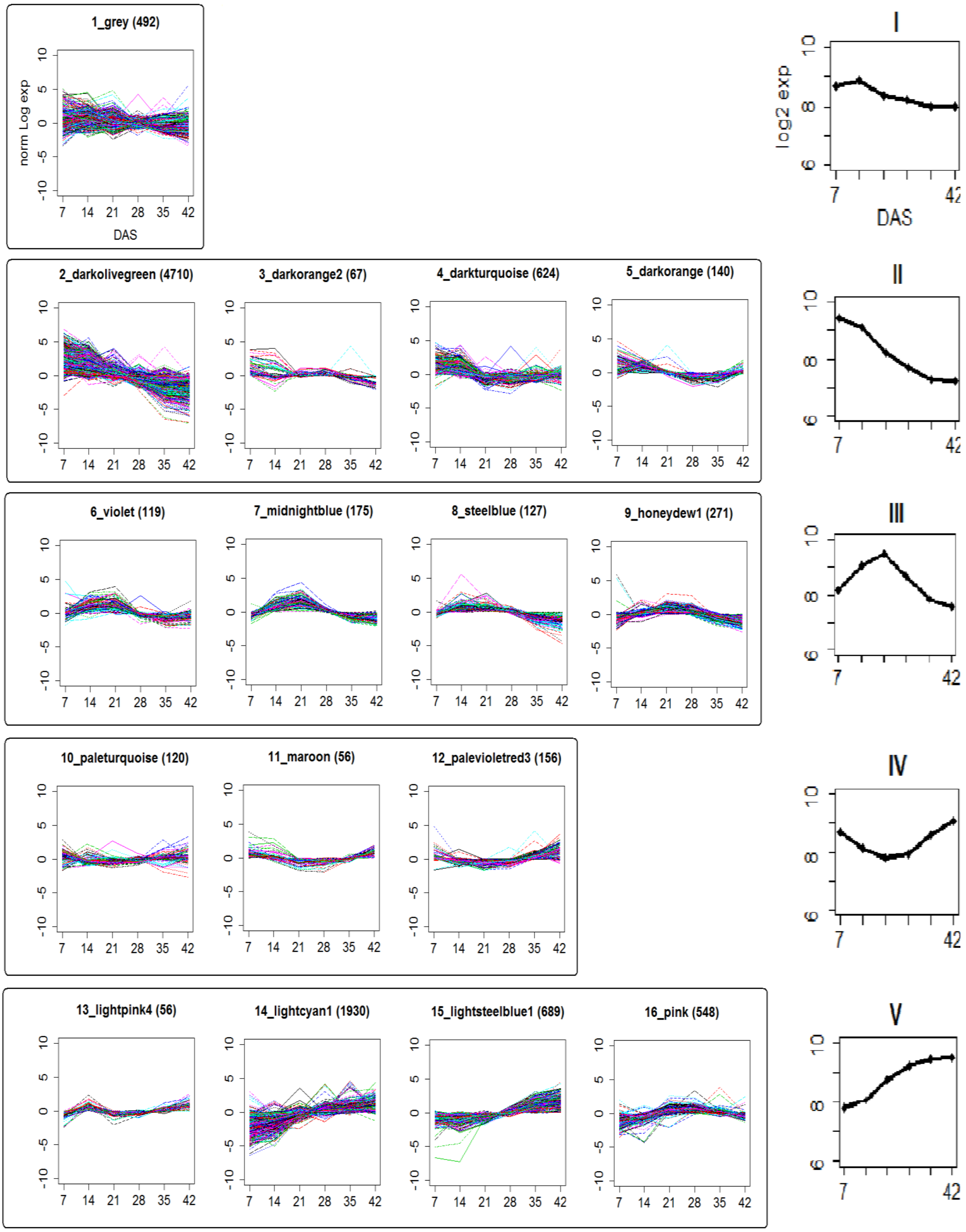
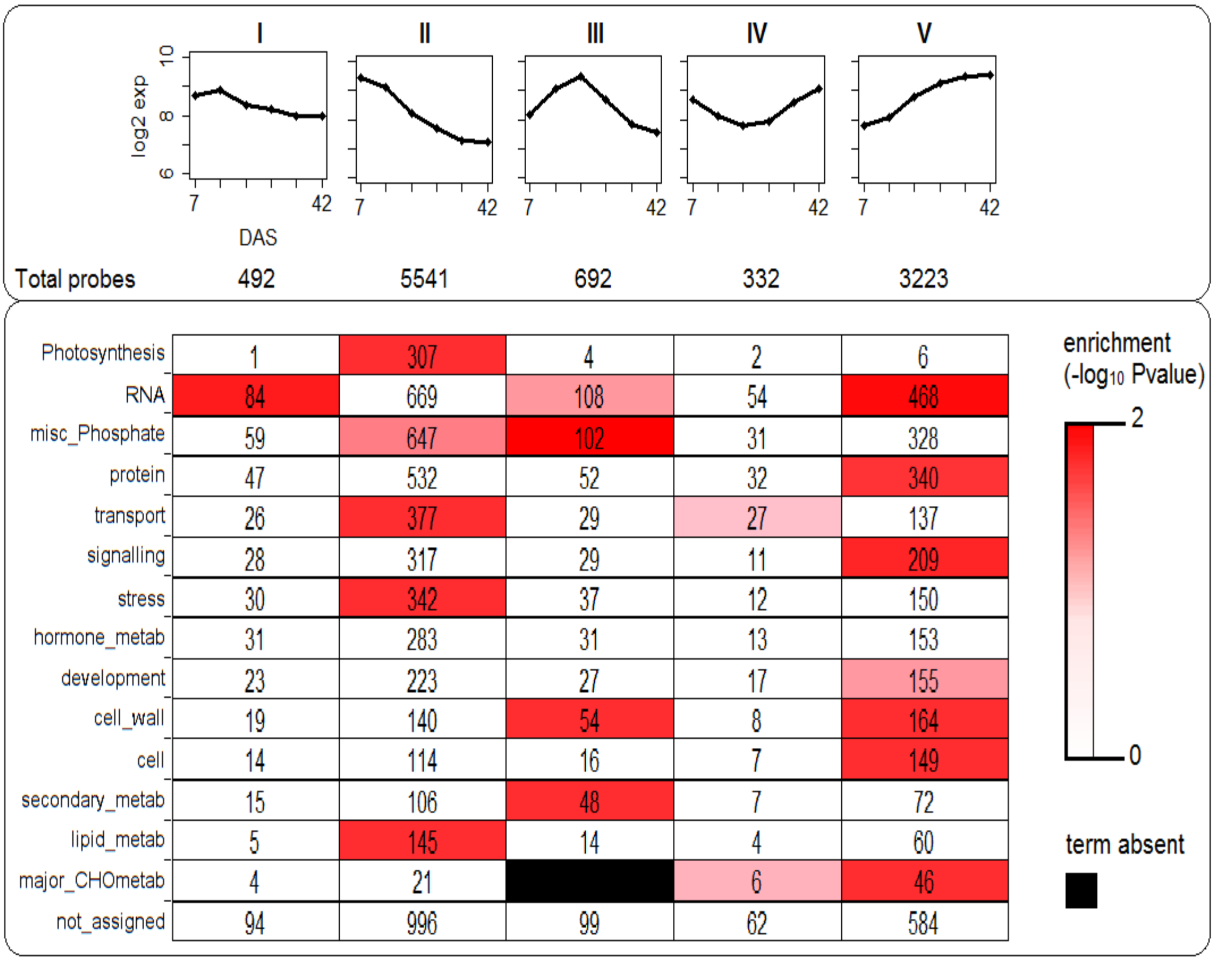


Figure 7.Co-expression modules resulted from WGCNA analysis named with a number and colours corresponding to the dendrogram in Appendix B *Figure 2*. The number of probes in each module is shown within brackets. The expression profiles are presented as expression normalized to the mean for each probe (y-axis) versus days after sowing (DAS, x-axis). The normalization by subtracting the mean allows to highlight the profiles rather than the magnitude of expression. The sixteen modules were further grouped in five clusters for pathway analysis (right column). Modules grouped on the same cluster are indicated by a square line and the resulting cluster is displayed next to that. Clusters are visualized by the mean expression profiles of all the probes.

Pathway analysis

In order to determine associations between different transcriptional profiles and biological functions of the genes, pathway analysis was performed aiming to detect significant overrepresentation of annotation terms in the modules identified through WGCNA. The five major clusters were used to perform pathway analysis. The MapMan ontology was used to map the probes to 35 functional categories (BIN). All the probes used for WGCNA could be mapped to a MapMan BIN. Probes designed on unique *B. rapa* genes that have no significant homology to *A. thaliana* genes, were assigned to the BIN 35 (“not_assigned”). Association between MapMan BINs and clusters was tested with Fisher`s exact test at p-value < 0.05 and the results for the fifteen BINs with higher number of annotated probes are shown in *Figure 8*. Results for the other twenty BINs are presented in Appendix C *Figure* 1. A significant overrepresentation could not be found for all of the functional categories. In fact, thirteen BINs (“hormone metabolism” , “not_assigned”, *Figure.8*; “redox”, “minor CHO metabolism”, “TCA”, “cofactor-vitamin metabolism”, “tetrapyrrole synthesis”, “glycolysis”, “biodegradation of xenobiotics”, “OPP”, “ATP synthesis”, “glyoxylate cycle”, “polyamine metabolism”, “S-assimilation”, Appendix C *Figure 1*) were not over-represented in any of the clusters.



	I	II	III	IV	V
Total probes	492	5541	692	332	3223
Photosynthesis	1	307	4	2	6
RNA	84	669	108	54	468
misc_Phosphate	59	647	102	31	328
protein	47	532	52	32	340
transport	26	377	29	27	137
signalling	28	317	29	11	209
stress	30	342	37	12	150
hormone_metab	31	283	31	13	153
development	23	223	27	17	155
cell_wall	19	140	54	8	164
cell	14	114	16	7	149
secondary_metab	15	106	48	7	72
lipid_metab	5	145	14	4	60
major_CHOmetab	4	21	■	6	46
not_assigned	94	996	99	62	584

Figure 8. Pathway analysis result for the fifteen MapMan BINs with the highest number of probes over five clusters. The top graphic shows the expression profiles for the five clusters and the total number of probes in each of the cluster. Expression profiles are represented as mean log 2 expression levels of all the probes (y-axis) over days after sowing (DAS, x-axis). The bottom graphic shows the number of annotated term of each BIN over the five clusters. Different shades of red are used to shows the significance of the Fisher`s exact text expressed as –log10(P-value). Significance threshold was set at –log10(0.05) = 1.3 and is shown with light red colour. Increasing colour intensity corresponds to lower P-values. Black spots indicate absences of that BIN in the relative cluster.

Conversely, six BINs were overrepresented in more than one cluster ("RNA", "misc_Phosphate", "transport", "cell wall", "major CHO metabolism"; *Figure 8;* "amino acid metabolism", Appendix C, *Figure 1*). The functional categories that showed enrichment were further investigated to study if the overrepresentation was determined by specific pathways (subcategories) that composed each BIN. As expected, most of the terms annotating specific pathways showed no significant enrichment. This can be explained with the size of the five clusters which resulted too big to be characterized with this analysis. Therefore, in the following paragraphs, specific pathways of each BIN are mentioned only when the majority of the probes were present in the cluster enriched for that specific BIN.

Photosynthesis related genes were strongly down-regulated during turnip development. In fact, almost all the probes mapping genes related to photosynthesis processes (307, 95.3%) are clustered in profile II showing a constant reduction in transcripts across the six time points. This is not surprising as the turnip tuber has a function as storage organ. Downregulation of genes associated with photosynthesis has been observed also in developing embryo in *B. rapa* (Basnet et al. 2013). Genes related to lipid metabolism are over-represented in cluster II with more than half of the annotated probes (145, 63.59%).These probes were mainly involved in fatty acids synthesis and elongation, lipid degradation and lipid transfer proteins. A third functional category with terms that enriched profile II was stress with 342 probes (59.89%), mainly represented by genes involved in plant defense mechanisms and pathogen –related (PR) proteins.

The BIN "misc_Phosphate" is a miscellanea group of enzymes involved in many different processes. It showed enrichment in profile II and III, meaning that the probes were highly expressed at earlier stages (7 – 14 DAS). This BIN includes Class III peroxidases proteins which are considered to play an important role during plant development due to their role in cell wall loosing and stiffening (Francoz et al. 2015).

Cluster III presented also enrichment for other three functional categories: cell wall, RNA and secondary metabolism. Cell wall counted 14% of the annotated probes (54 out of 385) and contained genes coding for UGP-dehydrogenases (UGD), enzymes involved in the synthesis of cell wall carbohydrates precursors, and cellulose synthases (CesA), which synthesize the glucan microfibrils in the apoplastic space (Olek et al. 2014). RNA BIN was represented by most of the transcription factors (TFs) families. Particularly, the MYB domain had many members showing a transient upregulated expression with peak at 21-28 DAS, during tuber initiation. MYB85, MYB46 and MYB63 were included in this cluster. Those three genes are considered upstream of the transcriptional network that regulates secondary cell wall biosynthesis (Schuetz et al. 2012). This correlated with the observation that turnip tuber at 28 DAS presented differentiate secondary growth elements (Zhang et al. 2014). Another TF family with members in cluster III was the AUX/IAA family, including TFs that regulate plant development and auxin-induced gene expression (Reed 2001).

Cluster V grouped probes with high transcript abundance at later stages of turnip tuber development. RNA BIN was enriched with 468 probes (33% of the total), mainly belonging to B3 and Triple-helix TFs families. Protein also showed enrichment in cluster V, however no specific subcategories were found overrepresented. Interestingly, genes associated to plant development and cell cycle showed enrichment only for this cluster. Especially, all the genes annotated for the subcategories cell cycle and cell division were upregulated throughout turnip tuber development. Similarly, upregulation of cell division genes was observed during development of tuberous sweet potato roots (Firon et al. 2013).

Overall, pathway analysis of co-expressed genes provided an overview of the main biological processes that turnip tuber underwent during development. The five clusters showed enrichment for multiple MapMan categories. This was expected considering the size of the clusters and the cross-talk between many pathways that may occur during plant development. On the other side, BINs certainly involved in the regulation of tuber development, for instance hormone metabolism, showed no enrichment for any of the clusters.

Overall, pathway analysis resulted limited for fully describing tuber development. Many MapMan BINs showed no enrichment in any of the clusters. On the other side, when BINs showed enrichment, the inspection of subcategories showed that few of them were overrepresented in single clusters. In order to characterize a specific pathway during turnip development, a different approach was then used.

Insights into sucrose and starch metabolism

The availability of a custom genome-wide microarray dataset allowed the parallel investigation of genes belonging to the same metabolic pathway during turnip tuber development. The tuber represents the storage organ of the turnip and it is characterized by high carbohydrates content, mainly composed in sucrose, glucose, fructose and starch (Temesgen 2012; Zhang et al. 2014). Therefore, in order to gain insights on the transcriptional regulation of genes involved in the sucrose and starch metabolism, the MapMan BINs “major CHO metabolism” and “glycolysis” were selected for further investigation. However, an exhaustive characterization of the regulation of starch biosynthetic pathway appeared not possible based on the pathway analysis results. In fact, only cluster IV (transient downregulation) and V (upregulation) were enriched for “major_CHO_metabolism”, while no overrepresentation was found for “glycolysis”. Moreover, many genes coding for the enzyme of the pathway could not be found in any of the clusters as the WGCNA was implemented on a small subset of the data (10’280 out of 61,461 probes). A different approach for selecting genes was then used. Differential expression analysis was carried out on the complete dataset resulting in 28,751 differentially expressed genes between consecutive time points at FDR<0.05 and absolute fold change of 1.5. Genes annotated in the MapMan BINs “major CHO metabolism” and “glycolysis” were selected as genes involved in the starch biosynthetic pathway. A schematic representation of the fourteen principal reaction of starch pathway is presented in *Figure 9*. *B .rapa* genes that were differentially expressed between time points and that are coding for the enzymes of the reactions in *Figure 9* are listed in *Table 1*. Most of the genes orthologs of the same *Arabidopsis* gene presented similar expression profiles, therefore one only the most representative profile was chosen and displayed in the figure in order to facilitate the visualization. If *B. rapa* orthologs of the same genes showed different profiles, both were selected for representation (reactions 1A, 1B, 7, 11).

Sucrose is synthesized in the source tissues as mature leaves and transported through the phloem to sink tissues as young leaves, roots and storage organs as tubers (Farrar 1996). Once in the sink tissues of the plant, sucrose is unloaded from the phloem and transported to the sink cells, where it can be localized in the apoplast, in the cytoplasm or in the vacuole. Sucrose is first cleaved into glucose or fructose by the enzyme invertase, which can also be localized in the apoplast (cell wall), cytoplasm or vacuole (Sturm and Tang 1999). In total, eleven genes coding for invertase enzymes showed significant changes between time points (reaction 1, *Table1*). All invertase located in the cell wall (1A) were downregulated during turnip development, although Bra036653, one of the two orthologous of BFRUCT3, showed a transient peak of expression at 21-28 DAS. Vacuolar invertase were represented are represented by two orthologs of ATBETAFRUCT4 which showed differential expression but with different profiles. In fact, Bra019749 presented a relatively stable expression while Bra026984 was strongly upregulated from 21DAS. Five cytosolic invertase orthologs of four different *A. thaliana* genes were expressed in a constant up or down-regulation manner. In summary, while most of invertase including the ones located at the cell wall were downregulated, two cytosolic invertase (Bra011567 and Bra034659) were strongly up-regulated from 28 DAS, the first time point were turnip tuber was visible.

Alternatively, after being transported from the phloem, sucrose can be degraded by sucrose synthase (SUS) in fructose and uridine-5`-diphosphate (UDP) glucose (reaction 2). In total five *B. rapa* genes, orthologs of SUS6, SUS3 and SUS1, code for sucrose synthase and showed significant changes in expression although not in a coordinate manner. Orthologs of SUS6 (Bra003845 and Bra 015995) and SUS3 (Bra036282) had a slightly increasing expression with a transient down regulation at 35 and 21 DAS respectively. Orthologs of SUS1 (Bra002332 and Bra006578) instead were strong up-regulated at 21 DAS. Overall, while most of invertase appeared down-regulated, sucrose synthase were mainly up-regulated. For most of these genes, the switch in expression took place between 21 and 28 DAS, when the turnip tuber initiation is most likely to occur. These expression profiles contrast with measurements of enzymatic activity carried out during tuber grown over 66 days (Gupta et al. 2001). In fact, while turnip tuber was gaining most of its biomass, it was detected an increasing activity of invertase and a decreasing of sucrose synthase. On the other hand, studies in potato showed a correlation between developmental changes on transcript abundance and enzyme level at early stages of tuberization (Appeldoorn et al. 1997; Kloosterman et al. 2005). The balance between the activity of invertase and sucrose synthase is considered fundamental to determine sink development and it appeared to be conserved among plant species (Koch 2004). High levels of invertase are thought to be important for sink tissue initiation and expansion, while sucrose synthase are more relevant at later stages during organ storage and maturation. Therefore the discrepancy between the results here presented and the work of Gupta et al. (2001) could be explained with difference in the plant material and timing of the measurements, as the activity enzymes was assayed at later stages.

The products of sucrose cleavage may undergo many different downstream reactions. UDP-glucose can enter in the cell wall metabolism or be used to re-synthesize sucrose (reaction 3 and 4). Interestingly, the two reactions are directed towards sucrose synthesis but, while genes coding or sucrose-phosphate synthase (SPS, reaction 4) were all up-regulated during turnip development, genes coding for sucrose phosphatase (SPP, reaction 3) were also expressed in a down-regulated manner. Alternatively, UDP-glucose can be phosphorylated by the UGPase in glucose-1-phosphate (reaction 5). Similarly, glucose and fructose are phosphorylated in glucose-6-phosphate and fructose-6-phosphate by hexokinase and fructokinase respectively (reaction 6 and 7). Interestingly, genes coding for these three enzyme shown similar expression profiles as up-regulation with switch at earlier stages (14-21DAS) or transient up-regulation with peak at 21DAS. It was also notable that the three genes coding for the hexokinase are all orthologs of *Arabidopsis* HKL1. Together with the aforementioned enzymes of reaction 1A and 1B, HKL1 orthologs represent then another example of differential expression within paralogs.

The reversible isomerisation between the hexose phosphates is catalyzed by the enzyme phosphoglucose isomerase (PGI, reaction 8) and phosphoglucomutase (PGM, reaction 9-10) (Keeling and Myers 2010). None of the *B. rapa* genes coding for the enzyme PGI were differentially expressed, therefore it is not possible to speculate about the conversion glucose-fructose. Alternatively, glucose-6-P is converted in glucose-1-P by the enzyme PGM directly in the cytosol (cytosolic PGM, reaction9) or, after being imported, into the amyloplast (plastidic PGM, reaction 10). Synthesis of starch is dependent upon import of glucose-6-P in the amyloplast and mutants for both PGM isoforms have shown to hamper starch accumulation in potato (Tauberger et al. 2000; Fernie et al. 2002). An interesting observation was that two genes coding for cytosolic PGM were differentially expressed and both up-regulated during turnip tuber development. This contrasts with research in potato where the cytosolic PGM was to found to have a relative stable expression during tuber development (Kloosterman et al. 2005). Therefore the different transcription regulation of PGM between potato and turnip may suggest different abundance of the respective enzymes. This could determine the difference in accumulation of starch or simple sugars as glucose, fructose between the two species. However this hypothesis is fairly simplistic as it does not take into account post-translation modifications, protein turnover and the actual enzymatic activity, all processes that play a fundamental role in the regulation of starch metabolism (Kötting et al. 2010).

Finally, glucose-1-P molecule is processed in the amyloplast by glucose-1-phosphate adenylyl transferase (AGPase, reaction 11), granular and soluble starch synthase (SS, reaction 12) to finally result in amylose, and the first component of starch. Additionally, branching enzyme as the 1,4-α-glucan branching enzyme (SBE, reaction 13) may activate to change the structure and folding of the starch complexes resulting in production of amylopectin (Keeling and Myers 2010). A total of fourteen genes were differentially expressed for these three reactions during turnip tuber development (*Table 1*). For most of these genes, the expression appeared up-regulated with a switch at 28 or 35 DAS. This appeared delayed compared to enzymes upstream in the pathway (reactions 1, 2, 6, 7, 9) and it correlated with the morphological observations of filling turnip tuber. Moreover, the expression of enzymes 11-13 resulted highly coordinated. Similar observations were made during tuber development in potato (Kloosterman et al. 2005). This may suggest that the transcriptional regulation of the starch biosynthesis pathway in turnip resembles the one in potato. Although appealing, this comparison if far from being conclusive. In fact, it should be pointed out that the orthology between the genes described in Kloosterman et al. 2005 and the turnip genes was not explored. The relevance of studying the sequence similarities relies on the evidence that enzymes involved in reactions 11-13 form heterotetrametric complex, together with other isoforms, and/or have particular localization (Ballicora et al. 2004). Therefore a complete overview should consider all the genes of the pathway.

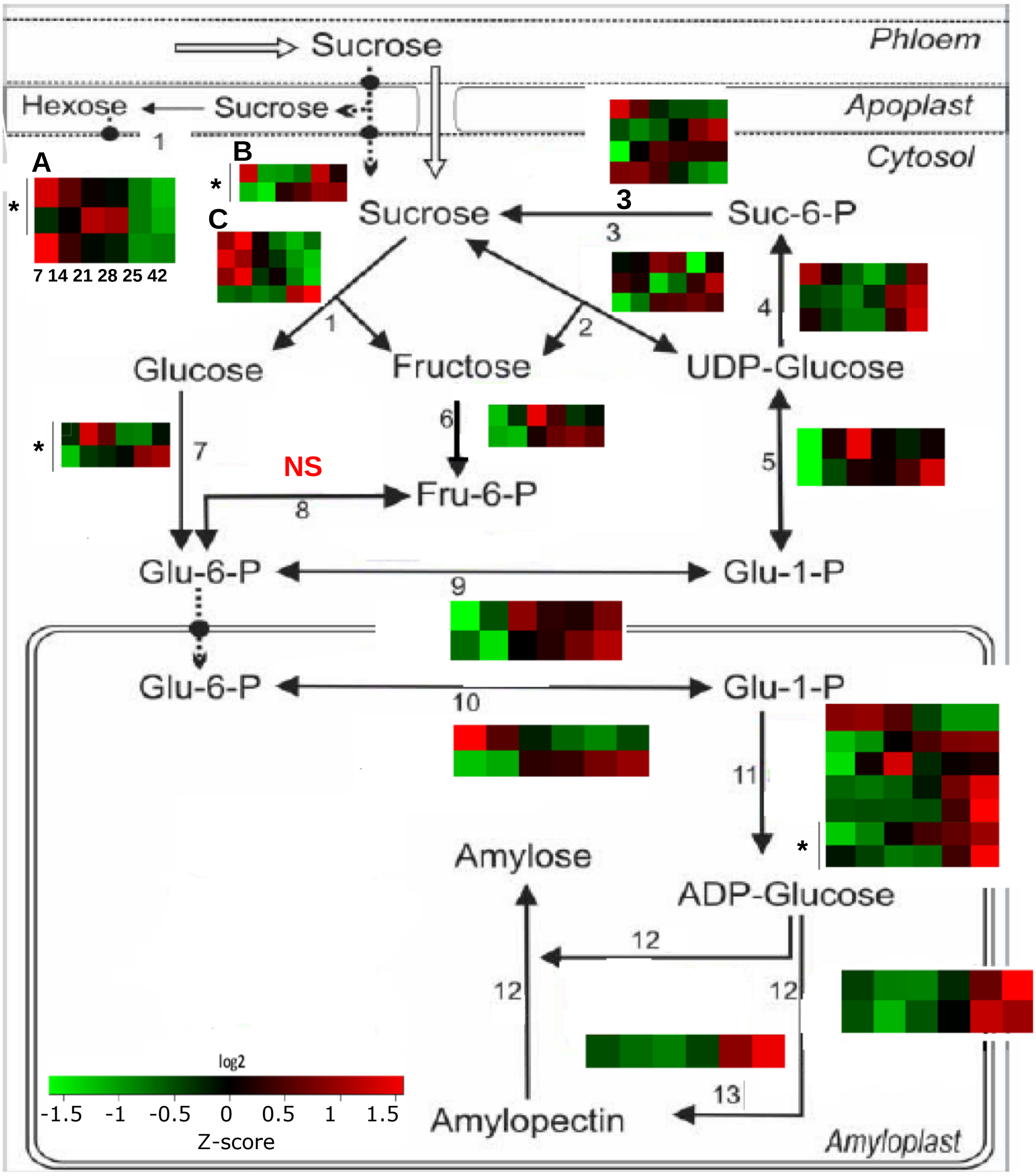


Figure 9. Schematic representation of the thirteen main reactions involved in the starch metabolism during six time points of turnip tuber development (7 – 42 DAS). Genes expression levels are represented as false colour boxes, one for each time point. Expression levels are represented by the Z-score which is obtained by normalizing the expression level to the mean and dividing by the standard deviation. Each reaction is summarized by the most representative profiles according to the orthology of *B. rapa* genes to *A .thaliana* gene. If all *B. rapa* genes orthologs to the same *A. thaliana* gene have similar profiles, only a representative profile is shown. Conversely, if orthologs to the same gene have different profiles, all of them are shown and indicated by a star (reactions, 1A, 1B, 7, 11). The reactions: 1A, cell wall invertase (orthologs to AT1G62660, AT3G13790); 1B, vacuolar invertase (AT1G12240); 1C cytsolic invertase (AT1G22650, AT1G35580, AT3G06500, AT4G34860); 2, sucrose synthase (AT1G73370, AT4G02280, AT5G20830); 3, sucrose phosphatase (AT1G51420, AT2G35840, AT3G54270); 4, sucrose phosphate-synthase (AT4G10120, AT5G11110, AT5G20280); 5, UGPase (AT5G17310, AT3G03250); 6, fructokinase (AT3G59480, AT2G31390, AT2G31390); 7, hexokinase (AT1G50460); 8, phosphoglucose isomerase (NS, no significant probes found); 9, cytosolic phosphoglucomutase (AT1G23190, AT5G51820); 10, plastidic phosphoglucomutase (AT1G70820, AT5G17530); 11, AGPase (AT5G19220, AT1G27680, AT1G74910, AT2G21590, AT2G21590, AT4G39210, AT5G48300); 12, starch synthase (AT1G32900, AT3G01180); 1,4-α-glucan branching enzyme (AT2G36390). A detailed list of all the significant *B. rapa* genes orthologs to the aforementioned *A. thaliana* gene is presented in *Table 1*. The background of the image is from Kloosterman et al. (2005), the expression profiles are from the original data of the present work.

Table 1. Classification of *B. rapa* genes putatively involved in starch metabolism.

KEGG annotation [international enzyme name]	***B. rapa* genes: differentially expressed / total annotated**	***B. genes* code**	**Homologous *Arabidopsis thaliana***	**reaction *Fig. 9***
cell wall invertase [EC 3.2.1.26]	4/15	Bra027030, Bra036653	AT1G62660 (BFRUCT3)	1a
		Bra021508, Bra027397	AT3G13790 (ATCWINV1)	
vacuolar invertase [EC 3.2.1.26]	2/3	Bra019749, Bra026984	AT1G12240 (ATBETAFRUCT4)	1b
cytosolic invertase [EC 3.2.1.26]	5/11	Bra016091	AT1G22650	1c
		Bra034413	AT1G35580 (CINV1)	
		Bra029583	AT3G06500	
		Bra011567, Bra034659	AT4G34860	
sucrose synthase (SUS) [EC 2.4.1.13]	5/7	Bra003845 Bra015995	AT1G73370 (SUS6)	2
		Bra036282	AT4G02280 (SUS3)	
		Bra002332, Bra006578	AT5G20830 (SUS1)	
sucrose-phosphatase (SPP) [EC 3.1.3.24]	5/7	Bra030439	AT1G51420 (SPP1)	3
		Bra023033, Bra017287	AT2G35840	
		Bra014826, Bra007060	AT3G54270 (SPP3)	
sucrose-phosphate synthase (SPS) [EC 2.4.1.14]	3/6	Bra033195	AT4G10120 (ATSPS4F)	4
		Bra006090	AT5G11110 (ATSPS2F, KNS2)	
		Bra002289	AT5G20280 (ATSPS1F)	
UTP-glucose-1-phosphate uridylyltransferase (UGPase) [EC 2.7.7.9]	2/2	Bra006395	AT5G17310 (UGP2)	5
		Bra032004	AT3G03250 (UGP1)	
fructokinase [EC 2.7.1.4]	4/15	Bra003378, Bra007452	AT3G59480	6
		Bra018248, Bra022839	AT2G31390	
hexokinase [EC 2.7.1.1]	3/10	Bra014254, Bra018850, Bra030490	AT1G50460 (HKL1)	7

Table 1. Classification of *B. rapa* genes putatively involved in starch metabolism.

KEGG annotation [international enzyme name]	***Brassica rapa* genes: differentially expressed / total annotated**	***B. genes* code**	**Homologous *Arabidopsis thaliana***	**reaction *Fig. 9***
cytosolic phosphoglucomutase (PGM) [EC 5.4.2.2]	2/3	Bra016357	AT1G23190 (PGM3)	9
		Bra028278	AT5G51820 (PGM1, STF1)	
plastidic phosphoglucomutase (PGM) [EC 5.4.2.2]	2/2	Bra016184	AT1G70820	10
		Bra023623	AT5G17530 (PGM2)	
glucose-1-phosphate adenylyltransferase (AGPase) [EC 2.7.7.27]	9/13	Bra002221, Bra023713	AT5G19220 (APL1)	11
		Bra032842	AT1G27680 (APL2)	
		Bra015883	AT1G74910	
		Bra026500, Bra030291	AT2G21590 (APL4)	
		Bra033604,	AT4G39210 (APL3)	
		Bra037495, Bra015135	AT5G48300 (ADG1)	
starch synthase (SS) [EC 2.4.1.21]	2/9	Bra010189	AT1G32900 (GBSS1)	12
		Bra021486	AT3G01180 (ATSS2)	
1,4-α-glucan branching enzyme (SBE)[EC 2.4.1.18]	1/3	Bra005269	AT2G36390 (BE3)	13

Conclusions

This work presented the results of the first genome-wide transcriptomic profiling during turnip tuber development. The availability of a custom *B. rapa* microarray represents a valuable resource to investigate transcriptional changes at the genome level. However, analysis of the duplicated probes on the microarray showed that the number of poorly correlated duplicated probes exceeds the number of *B. rapa* genes proven to undergo alternative splicing. This let us conclude that some probes may not detect the genes for which they have been designed, thus introducing a bias in downstream analysis.

A global overview of the transcript abundance showed that turnip tuber development requires massive changes in the all transcriptome. In fact, turnip tuber showed major morphological changes between 21 and 28 DAS when most of the genes showed changes in their expression profiles. The highest variation in transcript abundance occurred between 14 and 21 DAS, suggesting this time points as the right timing for further genetical genomics studies.

WGCNA identified 16 co-expressed modules that were organized with a scale free topology. In each module, few genes were highly connected and represented the hub genes, supposedly the key regulator genes of the module. Modules and hub genes can be further used for coexpression analysis.

The 16 modules were grouped in five clusters that were used for testing overrepresentation of MapMan functional categories. Results showed that the clusters were enriched for specific functional categories that characterize the different developmental stages. However, some pathways containing genes known to be involved in tuberization, as hormones, were not overrepresented. It can be conclude that this analysis only gives a broad overview of the biological processes involved in tuber development.

Turnip tubers accumulate sucrose, glucose, fructose but little starch. In order to obtain insights into the transcriptional regulation of the sucrose and starch metabolism, a detail analysis of the genes encoding the main enzymes of the pathway was carried out. Invertase and sucrose synthase showed opposite profiles, with invertase highly expressed at earlier stages and sucrose synthase at later stages. However, enzymatic assays on developing tuber of a different turnip accession showed the opposite behaviour (Gupta et al. 2001). Further research can investigate gene expression, enzymatic activity and sugar content on the same plant material to better characterize the balance between the two paths of sucrose cleavage paths.

Overall, most of the genes encoding enzymes downstream of sucrose synthase showed coordinate upregulation. This part of the pathway is also co-ordinately upregulated during potato tuber development. Considering this, comparison of expression profiles between the two species cannot elucidate their difference in starch accumulation. The only exception is represented by cytosolic phosphoglucomutase which appears strongly upregulated in turnip and relatively stable in potato. A variation in abundance and activity of this enzyme could influence the rate of glucose-6-phosphate imported into the amyloplast where the starch synthesis occurs. It would be interesting to validate this hypothesis by measuring carbohydrate content and gene expression among turnip accession varying for carbohydrates composition.

The characterization of sucrose and starch metabolism in turnip presented in this work is based on orthology of *B. rapa* genes with *A. thaliana* and selection of differential expressed probes across time points. However, not all the *A. thaliana* genes annotated in the pathway have been experimentally validated, therefore *B. rapa* genes that appeared differentially expressed may be not involved in the pathway. Moreover, processes as post-translation modifications, protein turnover and the actual enzymatic activity also play a role in determining the role of the different enzymes in a metabolic pathway. Considering this, the description of the sucrose and starch metabolism here presented is meant to provide a first insight into the transcriptional regulation but further research is recommended.

Appendix A – Microarray quality

Table 1. Correlation matrix displaying pairwise Pearson correlation coefficients (PCC)calculated on raw expression values of the twelve arrays across the six time points. Each array is labelled as “time point T of replicate R ”. PCC between biological replicates are highlighted by squares and present values above 0.99.

	T1_R1	T1_R2	T2_R1	T2_R2	T3_R1	T3_R2	T4_R1	T4_R2	T5_R1	T5_R2	T6_R1	T6_R2
T1_R1	1.000	0.991	0.977	0.975	0.944	0.936	0.916	0.917	0.886	0.886	0.880	0.878
T1_R2	0.991	1.000	0.972	0.970	0.936	0.930	0.911	0.910	0.882	0.881	0.875	0.873
T2_R1	0.977	0.972	1.000	0.995	0.969	0.961	0.941	0.942	0.909	0.911	0.904	0.903
T2_R2	0.975	0.970	0.995	1.000	0.970	0.962	0.941	0.942	0.909	0.911	0.905	0.903
T3_R1	0.944	0.936	0.969	0.970	1.000	0.994	0.980	0.979	0.953	0.953	0.945	0.942
T3_R2	0.936	0.930	0.961	0.962	0.994	1.000	0.983	0.981	0.959	0.956	0.946	0.943
T4_R1	0.916	0.911	0.941	0.941	0.980	0.983	1.000	0.995	0.980	0.979	0.969	0.964
T4_R2	0.917	0.910	0.942	0.942	0.979	0.981	0.995	1.000	0.982	0.984	0.974	0.970
T5_R1	0.886	0.882	0.909	0.909	0.953	0.959	0.980	0.982	1.000	0.991	0.984	0.982
T5_R2	0.886	0.881	0.911	0.911	0.953	0.956	0.979	0.984	0.991	1.000	0.991	0.989
T6_R1	0.880	0.875	0.904	0.905	0.945	0.946	0.969	0.974	0.984	0.991	1.000	0.995
T6_R2	0.878	0.873	0.903	0.903	0.942	0.943	0.964	0.970	0.982	0.989	0.995	1.000

Appendix B – WGCNA

Figure 1. Diagnostic plots for network construction. Left plot shows different β parameters (x-axis) relative to their fit of the free topology criterion expressed with R^2 (y-axis). The free topology criterion is set to R^2=0.8 (red line). A value of 18 was chosen for the β parameter as the first value below the red line. Right plot shows the distribution of the connectivity *k* (x-axis) of the resulting network for β = 18. The free scale topology is confirmed as few probes presented high connectivity being the hubs.

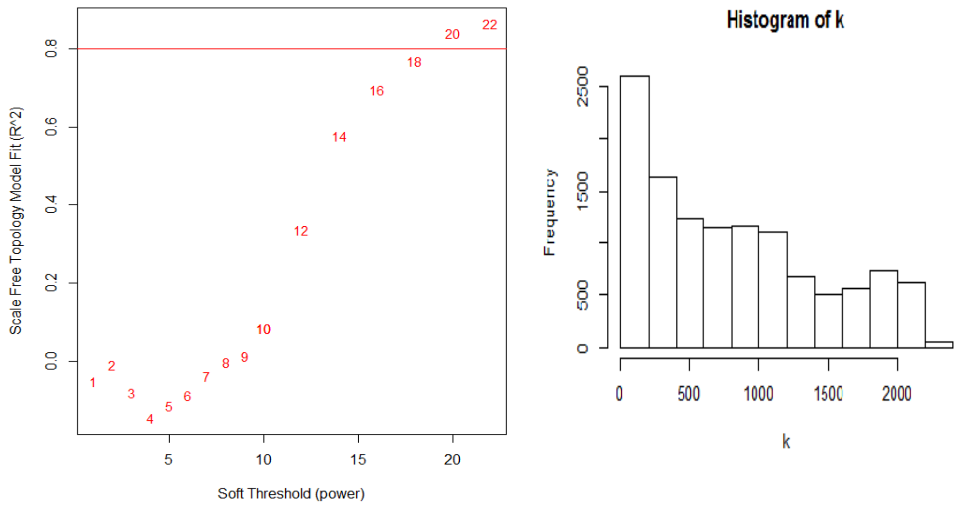


Figure 2. Construction of network with WGCNA. The dendrogram resulted from hierarchical clustering on a dissimilarity matrix base on DissTOM distance (see Methods). Branching cut of the dendrogram resulted in 56 co-expression modules (upper coloured bar) that were further merge if the correlation of MEs was higher than 0.9. This resulted in 16 final modules (lower coloured bar).

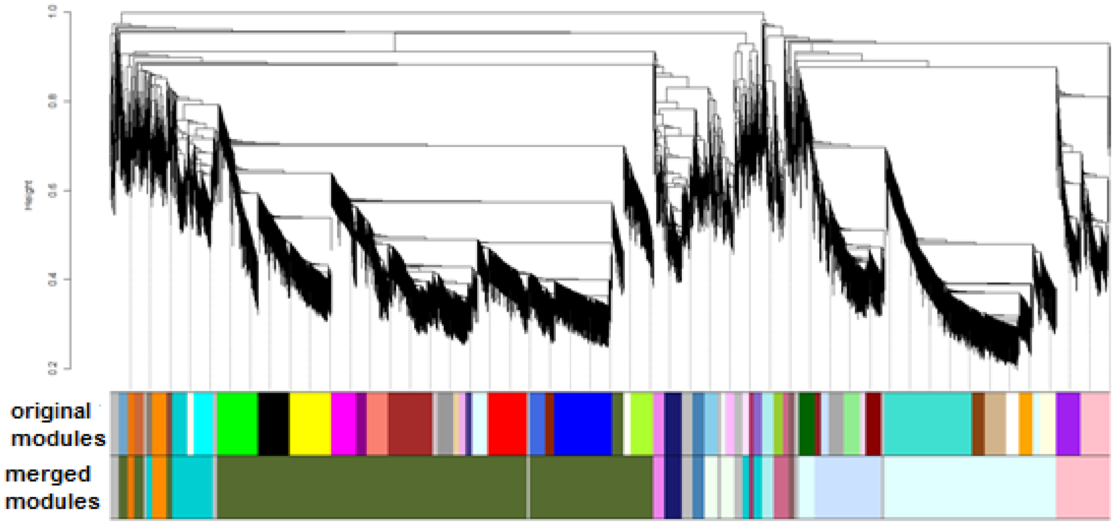
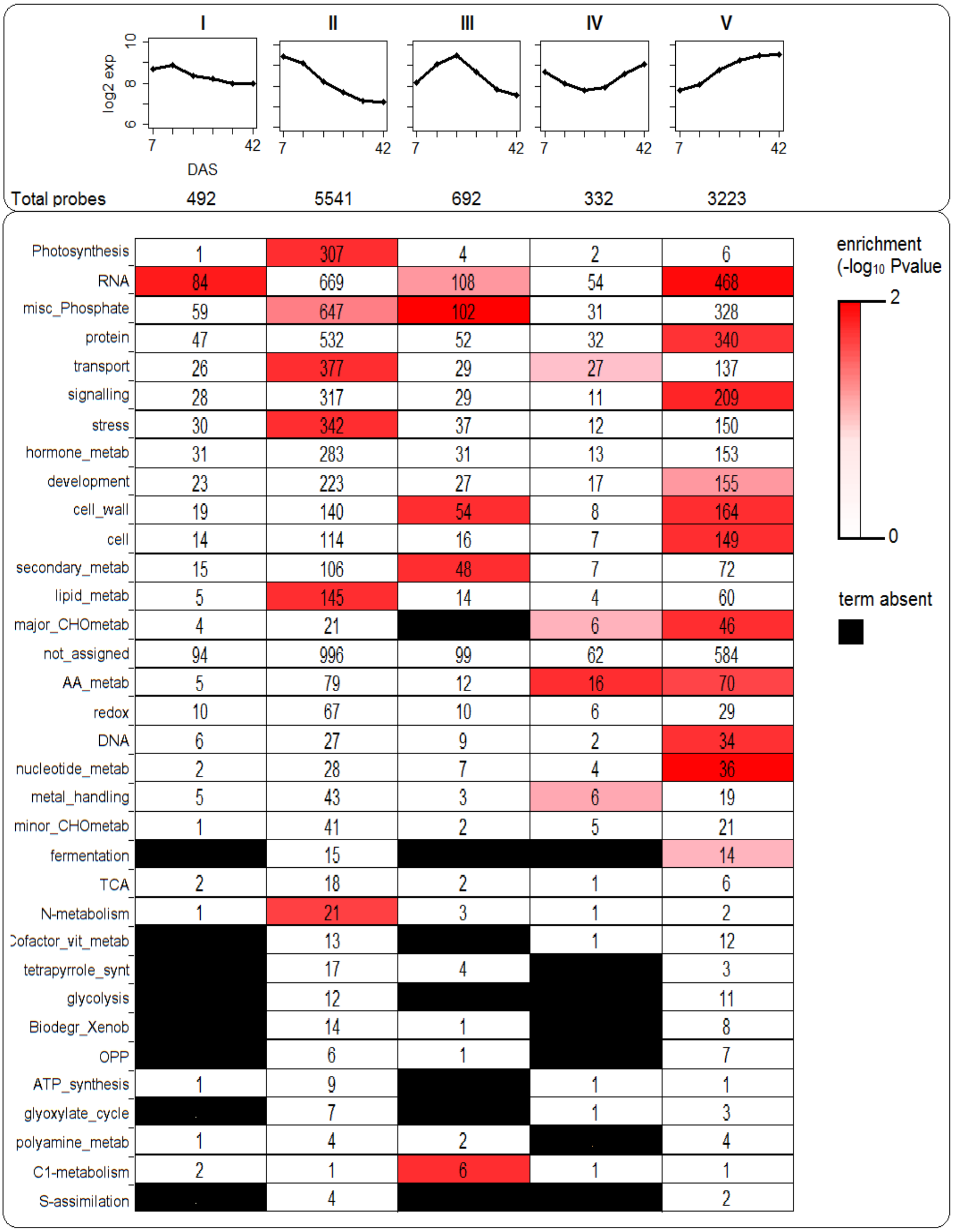


Table 1. Components resulted from Principal components analysis and variance explained.

variance	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
explained	42.6	16.4	8.7	6.1	4.8	4.5	4.2	3.5	3.2	2.8	2.9
cumulative	42.6	59.1	67.8	73.9	78.8	83.3	87.6	91.1	94.3	97.1	100

Appendix C – Enrichment analysis

Figure 1. Pathway analysis result for all the 35 MapMan BINs five clusters. The top graphic shows the expression profiles for the five clusters and the total number of probes in each of the cluster. Expression profiles are represented as mean log 2 expression levels of all the probes (y-axis) over days after sowing (DAS, x-axis). The bottom graphic shows the number of annotated term of each BIN over the five clusters. Different shades of red are used to shows the significance of the Fisher`s exact text expressed as –log10(P-value). Significance threshold was set at 0.05 (light red). Black spots indicate absences of that BIN in the relative cluster.



	I	II	III	IV	V
Photosynthesis	1	307	4	2	6
RNA	84	669	108	54	468
misc_Phosphate	59	647	102	31	328
protein	47	532	52	32	340
transport	26	377	29	27	137
signalling	28	317	29	11	209
stress	30	342	37	12	150
hormone_metab	31	283	31	13	153
development	23	223	27	17	155
cell_wall	19	140	54	8	164
cell	14	114	16	7	149
secondary_metab	15	106	48	7	72
lipid_metab	5	145	14	4	60
major_CHOmetab	4	21	■	6	46
not_assigned	94	996	99	62	584
AA_metab	5	79	12	16	70
redox	10	67	10	6	29
DNA	6	27	9	2	34
nucleotide_metab	2	28	7	4	36
metal_handling	5	43	3	6	19
minor_CHOmetab	1	41	2	5	21
fermentation	■	15	■	■	14
TCA	2	18	2	1	6
N-metabolism	1	21	3	1	2
Cofactor_vit_metab	■	13	■	1	12
tetrapyrrole_synt	■	17	4	■	3
glycolysis	■	12	■	■	11
Biodegr_Xenob	■	14	1	■	8
OPP	■	6	1	■	7
ATP_synthesis	1	9	■	1	1
glyoxylate_cycle	■	7	■	1	3
polyamine_metab	1	4	2	■	4
C1-metabolism	2	1	6	1	1
S-assimilation	■	4	■	■	2

References

Abelenda JA, Navarro C, Prat S (2014) Flowering and tuberization: a tale of two nightshades. Trends in plant science 19 (2):115-122

Appeldoorn NJ, de Bruijn SM, Koot-Gronsveld EA, Visser RG, Vreugdenhil D, van der Plas LH (1997) Developmental changes of enzymes involved in conversion of sucrose to hexose-phosphate during early tuberisation of potato. Planta 202 (2):220-226

Ballicora MA, Iglesias AA, Preiss J (2004) ADP-glucose pyrophosphorylase: a regulatory enzyme for plant starch synthesis. Photosynthesis research 79 (1):1-24

Basnet RK, Moreno-Pachon N, Lin K, Bucher J, Visser RG, Maliepaard C, Bonnema G (2013) Genome-wide analysis of coordinated transcript abundance during seed development in different Brassica rapa morphotypes. BMC genomics 14 (1):840

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

Bolstad BM, Irizarry RA, Åstrand M, Speed TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19 (2):185-193

Bonnema G, Carpio D, Zhao J (2011) Diversity analysis and molecular taxonomy of Brassica vegetable crops. Genetics, genomics and breeding of crop plants Enfield, USA: Science Publishers:81-124

Cheng F, Wu J, Wang X (2014) Genome triplication drove the diversification of Brassica plants. Horticulture Research 1

Cleveland WS (1979) Robust locally weighted regression and smoothing scatterplots. Journal of the American statistical association 74 (368):829-836

Del Carpio DP, Basnet RK, De Vos RC, Maliepaard C, Visser R, Bonnema G (2011) The patterns of population differentiation in a Brassica rapa core collection. Theoretical and Applied Genetics 122 (6):1105-1118

Deluc L, Grimplet J, Wheatley M, Tillett R, Quilici D, Osborne C, Schooley D, Schlauch K, Cushman J, Cramer G (2007) Transcriptomic and metabolite analyses of Cabernet Sauvignon grape berry development. BMC genomics 8 (1):429

Dixon GR (2007) Vegetable Brassicas and related crucifers. vol 14. CABI,

Dobbin K, Kawasaki ES, Petersen D, Simon R (2005) Characterizing dye bias in microarray experiments. Bioinformatics 21 (10):2430-2437

Fernie AR, Tauberger E, Lytovchenko A, Roessner U, Willmitzer L, Trethewey RN (2002) Antisense repression of cytosolic phosphoglucomutase in potato (Solanum tuberosum) results in severe growth retardation, reduction in tuber number and altered carbon metabolism. Planta 214 (4):510-520

Firon N, LaBonte D, Villordon A, Kfir Y, Solis J, Lapis E, Perlman TS, Doron-Faigenboim A, Hetzroni A, Althan L (2013) Transcriptional profiling of sweetpotato (Ipomoea batatas) roots indicates down-regulation of lignin biosynthesis and up-regulation of starch biosynthesis at an early stage of storage root formation. BMC genomics 14 (1):460

Francoz E, Ranocha P, Nguyen-Kim H, Jamet E, Burlat V, Dunand C (2015) Roles of cell wall peroxidases in plant development. Phytochemistry 112:15-21

Gupta AK, Singh J, Kaur N (2001) Sink development, sucrose metabolising enzymes and carbohydrate status in turnip (Brassica rapa L.). Acta Physiologiae Plantarum 23 (1):31-36

Habtemariam H (2012

) Gene expression study of turnip tuber formation. WUR MSc thesis

Horvath S, Dong J (2008) Geometric interpretation of gene coexpression network analysis. PLoS Comput Biol 4 (8):e1000117

Ihaka R, Gentleman R (1996) R: a language for data analysis and graphics. Journal of computational and graphical statistics 5 (3):299-314

Jackson SD (1999) Multiple signaling pathways control tuber induction in potato. Plant Physiology 119 (1):1-8

Jansen RC, Nap J-P (2001) Genetical genomics: the added value from segregation. TRENDS in Genetics 17 (7):388-391

Keeling PL, Myers AM (2010) Biochemistry and genetics of starch synthesis. Annual review of food science and technology 1:271-303

Kloosterman B, Abelenda JA, Gomez MdMC, Oortwijn M, de Boer JM, Kowitwanich K, Horvath BM, van Eck HJ, Smaczniak C, Prat S (2013) Naturally occurring allele diversity allows potato cultivation in northern latitudes. Nature 495 (7440):246-250

Kloosterman B, De Koeyer D, Griffiths R, Flinn B, Steuernagel B, Scholz U, Sonnewald S, Sonnewald U, Bryan GJ, Prat S (2008) Genes driving potato tuber initiation and growth: identification based on transcriptional changes using the POCI array. Functional & integrative genomics 8 (4):329-340

Kloosterman B, Vorst O, Hall RD, Visser RG, Bachem CW (2005) Tuber on a chip: differential gene expression during potato tuber development. Plant biotechnology journal 3 (5):505-519

Koch K (2004) Sucrose metabolism: regulatory mechanisms and pivotal roles in sugar sensing and plant development. Current opinion in plant biology 7 (3):235-246

Kötting O, Kossmann J, Zeeman SC, Lloyd JR (2010) Regulation of starch metabolism: the age of enlightenment? Current opinion in plant biology 13 (3):320-328

Kubo N, Saito M, Tsukazaki H, Kondo T, Matsumoto S, Hirai M (2010) Detection of quantitative trait loci controlling morphological traits in Brassica rapa L. Breeding science 60 (2):164-171

Langfelder P, Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. BMC bioinformatics 9 (1):559

Langfelder P, Horvath S (2012) Fast R functions for robust correlations and hierarchical clustering. Journal of statistical software 46 (11)

Langfelder P, Zhang B, Horvath S (2008) Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. Bioinformatics 24 (5):719-720

Lin K, Zhang N, Severing EI, Nijveen H, Cheng F, Visser RG, Wang X, de Ridder D, Bonnema G (2014) Beyond genomic variation-comparison and functional annotation of three Brassica rapa genomes: a turnip, a rapid cycling and a Chinese cabbage. BMC genomics 15 (1):250

Lou P, Zhao J, Kim JS, Shen S, Del Carpio DP, Song X, Jin M, Vreugdenhil D, Wang X, Koornneef M (2007) Quantitative trait loci for flowering time and morphological traits in multiple populations of Brassica rapa. Journal of Experimental Botany 58 (14):4005-4016

Lu G, Cao J, Yu X, Xiang X, Chen H (2008) Mapping QTLs for root morphological traits inBrassica rapa L. based on AFLP and RAPD markers. Journal of applied genetics 49 (1):23-31

Lukaszewska E, Virden R, Sliwinska E (2012) Hormonal control of endoreduplication in sugar beet (Beta vulgaris L.) seedlings growing in vitro. Plant Biology 14 (1):216-222

Mitsui Y, Shimomura M, Komatsu K, Namiki N, Shibata-Hatta M, Imai M, Katayose Y, Mukai Y, Kanamori H, Kurita K (2015) The radish genome and comprehensive gene expression profile of tuberous root formation and development. Scientific reports 5

Nishijima T, Sugii H, Fukino N, Mochizuki T (2005) Aerial tubers induced in turnip (Brassica rapa L. var. rapa (L.) Hartm.) by gibberellin treatment. Scientia horticulturae 105 (4):423-433

Olek AT, Rayon C, Makowski L, Kim HR, Ciesielski P, Badger J, Paul LN, Ghosh S, Kihara D, Crowley M (2014) The structure of the catalytic domain of a plant cellulose synthase and its assembly into dimers. The Plant Cell 26 (7):2996-3009

Peterson R (1973) Control of cambial activity in roots of turnip (Brassica rapa). Canadian Journal of Botany 51 (2):475-480

Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási A-L (2002) Hierarchical organization of modularity in metabolic networks. science 297 (5586):1551-1555

Reed JW (2001) Roles and activities of Aux/IAA proteins in Arabidopsis. Trends in plant science 6 (9):420-425

Rouhier H, Usuda H (2001) Spatial and temporal distribution of sucrose synthase in the radish hypocotyl in relation to thickening growth. Plant and Cell Physiology 42 (6):583-593

Roumeliotis E, Kloosterman B, Oortwijn M, Kohlen W, Bouwmeester HJ, Visser RG, Bachem CW (2012) The effects of auxin and strigolactones on tuber initiation and stolon architecture in potato. Journal of experimental botany 63 (12):4539-4547

Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, Vingron M, Schölkopf B, Weigel D, Lohmann JU (2005) A gene expression map of Arabidopsis thaliana development. Nature genetics 37 (5):501-506

Schuetz M, Smith R, Ellis B (2012) Xylem tissue specification, patterning, and differentiation mechanisms. Journal of experimental botany:ers287

Smyth GK (2005) Limma: linear models for microarray data. In: Bioinformatics and computational biology solutions using R and Bioconductor. Springer, pp 397-420

Smyth GK, Speed T (2003) Normalization of cDNA microarray data. Methods 31 (4):265-273

Sturm A, Tang G-Q (1999) The sucrose-cleaving enzymes of plants are crucial for development, growth and carbon partitioning. Trends in plant science 4 (10):401-407

Takahashi H, Kimura M, Suge H, Saito T (1994) Interactions between vernalization and photoperiod on the flowering and bolting of different turnip [Brassica rapa] varieties. Journal of the Japanese Society for Horticultural Science (Japan)

Tauberger E, Fernie AR, Emmermann M, Renz A, Kossmann J, Willmitzer L, Trethewey RN (2000) Antisense inhibition of plastidial phosphoglucomutase provides compelling evidence that potato tuber amyloplasts import carbon from the cytosol in the form of glucose-6-phosphate. The Plant Journal 23 (1):43-53

Temesgen M (2012) Investigating hormone regulation and sugar storage during turnip development in tunip plants. WUR MSc thesis

Thimm O, Bläsing O, Gibon Y, Nagel A, Meyer S, Krüger P, Selbig J, Müller LA, Rhee SY, Stitt M (2004) mapman: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. The Plant Journal 37 (6):914-939

Ting F-T, Wren M (1980) Storage organ development in radish (Raphanus sativus L.). 2. Effects of growth promoters on cambial activity in cultured roots, decapitated seedlings and intact plants. Annals of Botany 46 (3):277-284

Tipney H, Hunter L (2010) An introduction to effective use of enrichment analysis software. Human genomics 4 (3):202

Vreugdenhil D, Sergeeva LI (1999) Gibberellins and tuberization in potato. Potato Research 42 (3-4):471-481

Wan Y, Poole RL, Huttly AK, Toscano-Underwood C, Feeney K, Welham S, Gooding MJ, Mills C, Edwards KJ, Shewry PR (2008) Transcriptome analysis of grain development in hexaploid wheat. BMC genomics 9 (1):121

Wang X, Wang H, Wang J, Sun R, Wu J, Liu S, Bai Y, Mun J-H, Bancroft I, Cheng F (2011) The genome of the mesopolyploid crop species Brassica rapa. Nature genetics 43 (10):1035-1039

Wilcox RR (2012) Introduction to robust estimation and hypothesis testing. Academic Press,

Wilson L, Lowe S (1973) The anatomy of the root system in West Indian sweet potato (Ipomoea batatas (L.) Lam.) cultivars. Annals of Botany 37 (3):633-643

Xu X, Vreugdenhil D, van Lammeren AA (1998) Cell division and cell enlargement during potato tuber formation. Journal of Experimental Botany 49 (320):573-582

You MK, Hur CG, Ahn YS, Suh MC, Jeong BC, Shin JS, Bae JM (2003) Identification of genes possibly related to storage root induction in sweetpotato. FEBS letters 536 (1):101-105

Zahurak M, Parmigiani G, Yu W, Scharpf RB, Berman D, Schaeffer E, Shabbeer S, Cope L (2007) Pre-processing Agilent microarray data. BMC bioinformatics 8 (1):142

Zhang B, Horvath S (2005) A general framework for weighted gene co-expression network analysis. Statistical applications in genetics and molecular biology 4 (1)

Zhang N, Zhao J, Lens F, de Visser J, Menamo T, Fang W, Xiao D, Bucher J, Basnet RK, Lin K (2014) Morphology, Carbohydrate Composition and Vernalization Response in a Genetically Diverse Collection of Asian and European Turnips (Brassica rapa subsp. rapa). PloS one 9 (12):e114241

Zhao J, Wang X, Deng B, Lou P, Wu J, Sun R, Xu Z, Vromans J, Koornneef M, Bonnema G (2005) Genetic relationships within Brassica rapa as inferred from AFLP fingerprints. Theoretical and Applied Genetics 110 (7):1301-1314