Master's Thesis

Bioinformatics Analysis of Arabidopsis Thaliana Stem Cell Transcriptomes

---

Zohair Aashiq

First Examiner:	Prof. Dr. Rolf Backofen
Second Examiner:	Prof. Dr. Thomas Laux
Adviser:	Dr. Edwin Groot

Albert-Ludwigs-University Freiburg
Faculty of Engineering
Department of Computer Science
Chair for Bioinformatics
&
Institute of Biology III
May 28th, 2018

Writing period

15. 11. 2017 – 28. 05. 2018

Examiner

Prof. Dr. Rolf Backofen, Prof. Dr. Thomas Laux

Advisers

Dr. Edwin Groot

Declaration

I hereby declare, that I am the sole author and composer of my thesis and that no other sources or learning aids, other than those listed, have been used. Furthermore, I declare that I have acknowledged the work of others by providing detailed references of said work.
I hereby also declare, that my Thesis has not been prepared for another examination or assignment, either wholly or excerpts thereof.

Place, Date	Signature

Abbreviations

DNA	Deoxyribonucleic acid
cDNA	Complementary Deoxyribonucleic acid
RNA	Ribonucleic acid
QC	Quiescent center
CSC	Columella stem cell
CRC	Columella root cap
DEG	Differentially expressed gene
WOX5	WUSCHEL RELATED HOMEOBOX 5
mRNA	Messenger RNA
CDF4	Cyclic dof factor 4
GO	Gene Ontology

Abstract

In the root meristem of Arabidopsis thaliana, stem cells divide asymmetrically for the self renewal and generation of new cells for the root development. The quiescent centre (QC) cells are embedded in the stem cells. These QC cells manage a fine balance to maintain the individual characteristics of the surrounding cells and repress their differentiation. The WUSCHEL HOMEOBOX 5 (WOX5) is one of the master regulators of distal root stem cells. The WOX5 protein is expressed in the QC and moves from to the columella stem cells to keep them undifferentiated. However, the target genes of WOX5 are still unknown. We show the direct targets of WOX5, identify the trancriptomic signature of QC, CSC and CRC and identify the relationship of these cell types to WOX5 in this thesis.

Zusammenfassung

Die Entwicklung der Wurzel hängt von Stammzellen im Wurzelmeristem von Arabidopsis thalina ab, die sich asymmetrisch teilen und dadurch sowohl zur Erhaltung der Stammzell-Nische, als auch zur Produktion von neuen Zellen beitragen. In der Mitte der Stammzellen liegen organisierende Zellen (QC), dies sich selten teilen. Diese Zellen schaffen eine Balance zwischen den individuellen Eigenschaften der sie umgebenden Zellen und unterdrücken deren Differenzierung. Der Stammzellregulator WUSCHEL HOMEOBOX RELATED 5 (WOX5), ist im QC exprimiert und bewegt sich von da zu den Columella-Stammzellen und hält sie undifferenziert. Die Ziel-Gene von WOX5 sind noch nicht bekannt. In dieser Arbeit zeigen wir direkte Ziel-Gene von WOX5, beschreiben die transkriptionelle Identität von QC, CSC und CRC und untersuchen den Zusammenhang zwischen WOX5 Expression und diesen Zelltypen.

Contents

List of Figures

List of Tables

1 Introduction

In plants, stem cells are responsible for the development and generation of tissues and other organs. These cells are generally defined by their function of self-renewal and generation of daughter cells [5]. The stem cells are embedded in the root and shoot apical meristems which are specialized tissues located on the tip of each side of the plant body [2]. The shoot meristem produces all aerial organs and tissues of the plant. On the other hand, the primary root of the plant arises from the stem cells located in the root meristems. A pool of stem cells surround a small group of organizing cells, the quiescent center cells (QCs) [6]. The quiescent center cells (QCs), columella stem cells (CSCs) and columella root cap cells (CRCs) together form a highly regulated environment called stem cell niche [7]. The columella stem cell niche is the smallest part of the root stem cell niche (see Figure1). The quiescent center cells are the organizer of stem cell niche and signals from organizer cells act to prevent stem cell differentiation [8]. These signals need the specific expression of the transcription factor WUSCHEL RELATED HOMEOBOX 5 (WOX5) in the QC [9].

The WOX5 are the WUS family of homeodomain transcription factors which are expressed in QC and moves to underlying cells in Arabidopsis thaliana. This transcription factor directly represses the differentiation factor CYCLING DOF FACTOR 4 (CDF4) which plays a crucial role in maintaining the differentiated state of the stem cells. However, knockdown of CDF4 does not restore stem cell loss in the WOX5 mutant which shows that WOX5 also binds to some more targets [1].

The additional targets of the WOX5 and interactions of genes expression involved in determining stem cell niche fate are unknown. So far, only mutant analysis of a few genes provided the information on the specification of stem cell niche cell types. The objective of this thesis is to develop the methods to identify and rank the transcriptomic signature of these cell types and their relation to WOX5 targets.

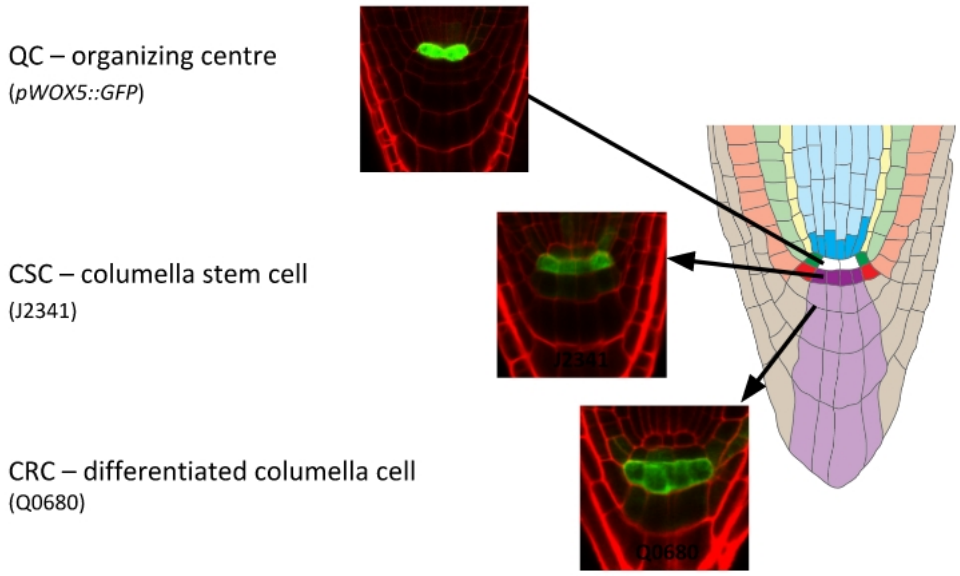


Figure 1: The stem cell niche concept (modified from Pi et al.2015 [1] [2]). Transcriptional reporter pWOX5::GFP, J2341 used for QC, CSC and CRC respectively.

Methods of linear model, moderated t-statistics, gene ontology analysis and biclustering are used for identifying and ranking of the transcriptomic signature of QC, CSC and CRC. "Bayesian Modeling of ChIP-chip Data Through a High-Order Ising Model" combined with differential expression analysis are used for finding the direct targets of WOX5.

This thesis deals with topics in biology, data mining and data analysis. It is divided into seven chapters. Chapter 2 shows the idea behind the bioiformatics methods, algorithms and gives the description about these methods. Chapter 3 discusses the related work. Chapter 4 illustrates the implementation of algorithms and methods for achieving the objective of this thesis. Chapter 5 shows the results of these methods and briefly discusses their interpretation. Chapter 6 discusses the results and finally chapter 7 gives the suggestions for the future work. People who are interested in repeating the experiments which were made as part of this thesis can contact the author via e-mail, to get the relevant data for the evaluation, scripts or source code

2 Background

All cells in an organism have same genomic DNA. The functions of the cells differ because of difference in gene expressions. While an organism contains a huge number of genes, just a few of them are actively expressed as mRNAs at a time because if all genes are expressed at the same time, there would be a lot of useless stream of structured information. Similarly, specific genes are expressed for performing different functions like responding to internal and external stimuli, resulting in growth, development and survival etc. The genes are processed to form mRNA which is used as a template for producing the proteins. These mRNAs expressed at any given time are called transcriptomes. The genes also interact with each other in the form of a complex network. The activation of one gene can lead to changes in expression of multiple genes. Understanding the connection between different genes is difficult because in molecular biology one can only work on a single gene at a time. This problem can be solved by bioinformatics analysis.

2.1 DNA Microarray

The DNA microarray is one of the technologies which is used to detect and analyzed thousands of genes in a sample [10]. This allows biologists to profile and study the transcriptomes. The microarray is a glass slid with a matrix of spots print that allows the quantification of mRNA transcripts which are present in the cells. The basic technology of microarray is based on hybridization of complementary sequence. Every spot on the chip represents a different coding sequence from different genes and each spot on the chip is made of a DNA probe that can pair with the complementary DNA (cDNA). There are four major steps of a microarray experiment (shown in Figure 2).

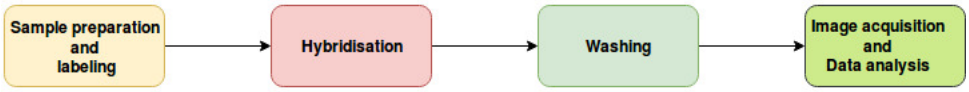


Figure 2: Steps of microarray experiment

To start an experiment, a "total RNA" containing mRNA is isolated from the subjected cell that ideally represents a quantitative copy of genes expressed at the time of sample collection [11]. The complementary DNA is prepared with degraded mRNA using the reverse-transcriptase enzyme. Each DNA is labelled with fluorescent dyes and the resulting labelled transcripts are called targets (shown in Figure 3). The samples are applied to the microarray and left in the hybridization chamber for a few hours. The cDNA binds to complementary base pairs in each of the spots on the array. Furthermore, the targets which are not hybridized are eliminated by washing them. The previous steps are different for two types of chips (shown in Figure 3).

- Spotted microarrays are called "two-color arrays" because the cDNA from two different cells of interest, labelled with fluorescent dyes of different colors are hybridized to a single chip.

- One color chip requires more slides per experiment because the tissue of interest is hybridized to one sample per chip.

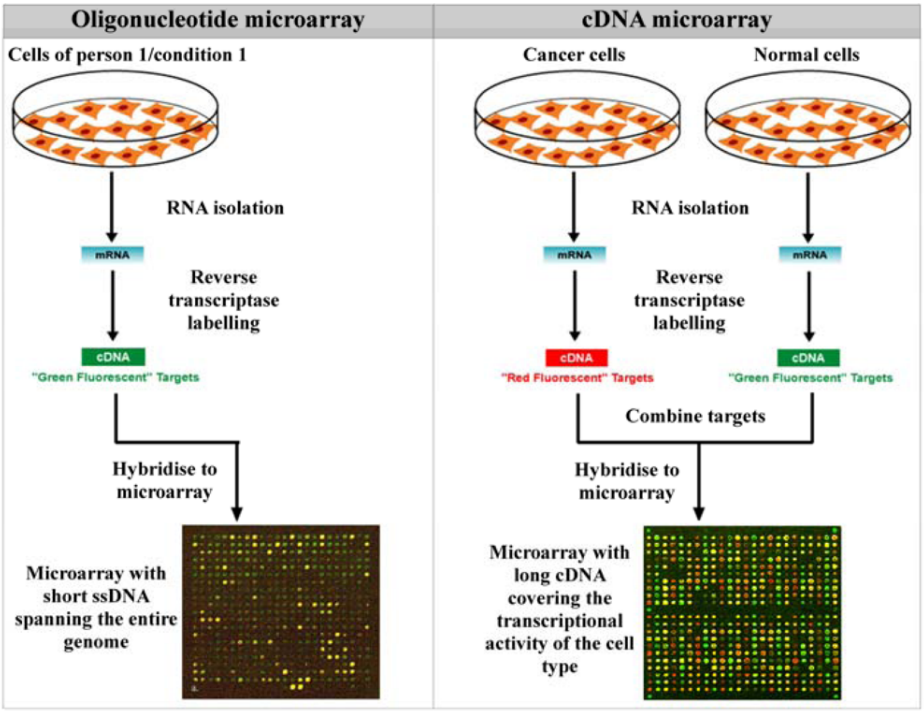


Figure 3: Workflow summary of one channel array and two channel microarrays. The image shows the RNA extraction, cDNA production, fluorescent labelling and hybridization of labelled target [3].

After slides are dried, they are scanned to determine the amount of cDNA bound to a certain quantity of labelled target. Consider a probe for a gene, the probe becomes red, if there is more mRNA in target cells. If there is more mRNA in reference cells, the probe becomes green. If both cells have the same amount of mRNA then the probe becomes yellow. The images of red and green fluorescence are taken separately using the laser and scanner. The scanner is used to capture the emission of fluorescence of the hybridized sample caused by laser light. The captured fluorescence turns into an image. Te image is then converted into numbers for further analysis. Image processing software is used for measuring the intensities.

Image processing software generates many quantities for each spot. It generates (i) Spot boundaries (ii) Foreground intensities (iii) Background intensities. The median red signal of foreground for each probe is denoted by R at each spot and the median of the green signal for each probe is denoted by G. The background intensities are local fluorescence surrounding the spot of hybridization. These are usually denoted by R_b and G_b. The background signals are subtracted from foreground signals to measure the fluorescence of hybridization. These intensities are used to yield the expressions which are scanned to determine the amount of cDNA bound to a certain quantity of labelled target. Consider a probe for a gene where M denotes the ratio between R and G:

$$M = \frac{R}{G} \tag{1}$$

The background-corrected expression ratio is given by:

$$M = \frac{R - R_b}{G - G_b} \tag{2}$$

In case of single microarray, there are single colored intensities at each probe. This array consists of two types of probe cells, the perfect match (PM) and the mismatch (MM). The perfect match (PM) represents the real expression and the mismatch (MM) represents the nonspecific hybridization signal in overall intensity measure. The formula for naive estimation is given by:

$$avg.diff = \frac{1}{|A|} \sum_{j \in A} (PM_j - MM_j) \tag{3}$$

where A is a set of probe pairs whose intensities do not deviate more than three times the standard deviation of the mean intensity over all probes [12].

2.1.1 Error correction and quality control

The required number of steps for analyzing the microarray data are shown in Figure 4. The main components of this analysis are experimental design, quality control, prepossessing and statistical analysis.

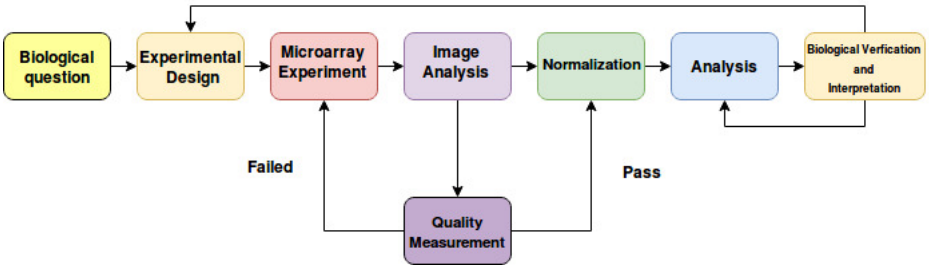


Figure 4: The Microarray Analysis Process

2.1.2 Quality control

The aim of the quality control is to make reliable data for further analysis. The number of quality control steps are as follows:

Array selection

There may be obvious flaws in arrays which are observable by visual inspection. By removing those arrays, results may improve.

Spot filtering

The results may also be improved by elimination of some specific spots suggested by visual identification of defects of printing or washing.

2.1.3 Data prepossessing

The data of microarray can have strong biases which must be corrected. There are two types of biases. The first type of bias is due to probe effect, in which probes of outside the binding region have higher intensities than those inside the bounding region probes. This type of bias can be corrected by background correction.

The other type of bias is due to difference in scanner settings, room temperatures and technician's expertise levels, etc. This bias may generate the variation of

probe distribution across the arrays. This type of bias could be corrected with the normalization methods by matching the distribution of probe intensities.

Background correction

One of the problems with image analysis is that some background signals can arise from non-specific binding and some of them may be due to non-biological source.

The goal of background correction is to remove the effects of non-binding or spatial heterogeneity. The background correction adjusts the foreground intensity to background fluorescence and gets a signal value which is proportional to expression level.

The commonly used methods for adjusting the background are normexp, subtract, edwards and movingmin. Normexp [13] method uses convolution model for background correction. Edwards [14] method is a log-linear interpolation method which adjusts lower intensities. Movingmin method replaces the background estimates with the minimums of the backgrounds of the spot and its eight neighbours. Subtract method is the most simple method for background correction in which the background intensity (R_b,G_b) is subtracted from foreground intensity (R_f,G_f) to give the true signal from hybridization ($R = R_f - R_b$, $G = G_f - G_b$).

Normalization

The main goal of the microarray data analysis is to identify the expression changes between the sample and technical variation which may lead to incorrect results. It is very important to identify these biases.

Normalization is an essential step to remove the systematic biases caused by different incorporation of dyes, different amounts of mRNA and different scanning parameters which may affect the measured gene expression levels. The goal of the normalization is to give the same distribution of each array for making multiple arrays comparable.

As the first step, one needs to decide a normalization method. There are a number of different normalization methods which are developed and tested for spotted microarrays [15][16]. Symth et al. [17] described some most commonly used methods. As one method cannot correct all technical artifacts, different methods were developed over period of time. There are two broad classes of normalization methods.

- Within-Array Normalization.

- Between-Array Normalization.

Within-Array normalization normalizes the M-values for one or more two-color spotted microarrays. These are usually used for traditional log-ratios of two-color data. Within-array normalization is usually not relevant for single-channel arrays.

Between-Array normalization is the sole normalization step for single-channel arrays. After background correction, the data is generally transformed into log-ratios. The log transformation improves the characteristics of the data distribution and allows the use of classical parametric statistical analysis. For single-channel data, quantile, scale or cyclicloess normalization methods work better. When an object is a matrix or EListRaw object, other methods produce an error. The quantile and cyclicloess normalization were proposed by Bolstad et al. [18]. The quantile, scale or cyclicloess normalization methods are called complete methods because they make use of data from all arrays in an experiment to form the normalizing relation. Cyclicloess uses M vs A plot to analyze expression data, M is the difference in log expression values and A is the average of log expression values [18].

2.1.4 Differential Expression Analysis

The reason for performing the preprocessing steps is preparation of data for further statistical analysis. The purpose of any statistical analysis is to determine the outcome which is a key point in data analysis. Linear models are among the most used statistical methods for data analysis and are effectively used for microarray data [19]. These models define a linear relationship between observed values and experimental conditions. This allows very general experiments to be analyzed nearly as easily as a simple replicated experiment [20].

For microarray analysis, linear models compare two groups or multifactorial designs e.g genotype or treatment. This comparison is used for selecting significantly different genes between the conditions. Linear models can summarize the log-ratio of each gene for designed microarray experiment with some level of replication before testing for differential expression [19].

Symth et al. [19] considered the problem of identifying the differentially expressed genes in designed microarray. They reset the hierarchical model of Lonnstedt and Speed [21] in the context of general linear models. Symth et al. [19] used empirical Bayes method to borrow the information of genes to estimate the gene-specific variance and used posterior variances in the classical t-test. The proposed model is completely data dependent and estimate the hyperparameres by empirical Bayes approach [19].

The equation for linear model is

$$y = \boldsymbol{X}\beta + \epsilon \tag{4}$$

Linear model can be obtained for each gene

$$E(\boldsymbol{y_g}) = X\alpha_g \quad \text{and} \quad Var(\boldsymbol{y_g}) = W_g\sigma_g^2 \tag{5}$$

Where y is vector of expression data. $\boldsymbol{X}$ is the design matrix and α is a vector coefficient. The contrast matrix specifies the RNA targets used on array. The contrasts of interest are given by $\beta_j = C^T\ \alpha_j$ where C is the contrast matrix which specifies the comparison of interest. The coefficient component of the fitted model gives estimated values for the α_j and inner product $C^T\ \alpha_j$ gives estimated values for the β_j.

We assume that linear model is fitted to the responses for each gene to obtain coefficient estimators $\hat{\alpha_g}$, estimator s_g^2 of σ_g^2 and estimated covariance matrices.

$$Var(\hat{\alpha_g}) = V_g\sigma_g^2 \tag{6}$$

Where V_g is a positive definite matrix not depending on s_g^2. The contrast estimators are $\hat{\beta}_g = C^T\alpha_g$ with estimated covariance matrices.

$$Var(\hat{\beta}_g) = C^TV_gC\sigma_g^2 \tag{7}$$

The responses y_g are not necessarily assumed to be normal and the fitting of the linear model is not assumed to be by least squares. However the contrast estimators are assumed to be approximately normal with mean β_g and covariance matrix $C^TV_gC\sigma_g^2$ and the residual variances s_g^2 are assumed to follow approximately a scaled chi-square distribution. The unscaled covarinace matrix V_g may depend on α_g, for example if the robust regression is used to fit the linear model. If so, the covairance matrix is assumed to be evaluated at α_g and the dependence is assumed to be such that it can be ignored to a first order approximation. Let v_{gj} be the jth diagonal element of $Var(\hat{\beta}_g) = C^TV_gC$. The distribution assumption can be summarized by

$$\hat{\beta_{gi}}|\beta_{gj},\sigma_g^2 \sim \mathrm{N}(\beta_{gj}, v_{gj}\sigma_g^2) \tag{8}$$

$$s_g^2|\sigma_g^2 \sim \frac{\sigma_g^2}{d_g}\chi d_g \tag{9}$$

Hypothesis testing is

$$H_0 : \beta_{gj} = 0 \quad \text{vs} \quad H_1 : \beta_{gj} \neq 0 \tag{10}$$

Extend empirical Bayes method from Loonnstedt and Speed [21] to more general experiment. Hierarchical model for variances where variance prior is

$$\frac{1}{\sigma_g^2} \sim \frac{1}{d_0 s_0^2} \chi^2 d_0 \tag{11}$$

This distributional result assumes d_0 and s_0 to be given values. In practice, they need to be estimated from the data. This describes how the variances are expected to vary across genes. For any given j, we assume that a β_{gj} is non-zero with known probability

$$P(\beta_{gj} \neq 0) = p_j \tag{12}$$

$$\beta_{gj} | \sigma_g^2, \beta_{gj} \neq 0 \sim \mathrm{N}(0, v_{0j} \sigma_g^2) \tag{13}$$

The posterior values shrink the observed variances towards the prior values with the degree of shrinkage depending on the relative sizes of the observed and prior degrees of freedom. Posterior variance estimator is

$$\tilde{s_g}^2 = \frac{1}{E(\sigma_g^2 | s_g^2)} = \frac{d_0 s_0^2 + d_g s_g^2}{d_0 + d_g} \tag{14}$$

The moderated t-statistics

$$|t_{gj}| = \frac{|\hat{\beta_{gj}}|}{\tilde{s_g} \sqrt{v_{gj}}} \sim t_{d_o} + d_g, under H_0 \tag{15}$$

2.1.5 Biclustering

One of the objectives of this thesis is to identify the cell-specific genes of different cell types. Different cell types have different patterns of gene expression. It means that if the cell types are more different than gene expression patterns will be more different. One way to find these patterns is through clustering. Clustering is unsupervised machine learning technique which identifies groups of similar patterns. Machine

learning is a science which uses statistical methods to parse the data, learn it and make a prediction.
There are two broad categories of machine learning:

- *Supervised machine learning*, machine learns to predict from labelled input.
- *Unsupervised machine learning*, identify the internal representation of input e.g. clusters.

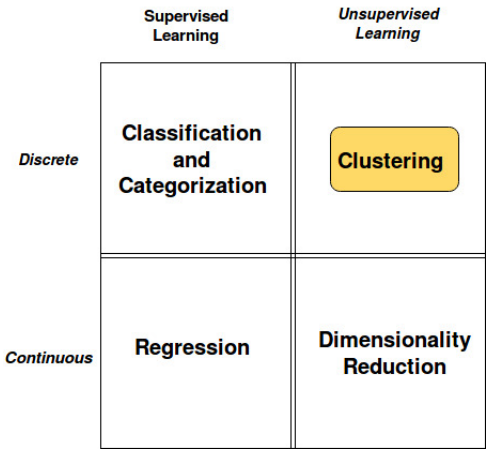


Figure 5: Machine learning problems

There are two other categories of machine learning depending on the output type:

- Discrete: Machine learning methods which give discrete output
- Continuous: Machine learning methods which give continuous output

Clustering is an unsupervised learning technique which finds the genes whose expressions fit specific predefined pattern in several biological and medical studies dealing with microarray data. The biological significance of clustering and its potential for microarray was discovered in early paper by Eisen et al. [22].

A pattern or cluster is defined as a set of objects which are similar to each other and dissimilar to other objects. In genomics data, the rows represent the genes and columns represent the measurement from different conditions. Each point characterizes the expression level of gene expression.

There are two basic types of methods for clustering (shown in Figure 6).

- Hierarchical method generates clusters by organizing the data into dendrogram or tree
- Partitioning method partitions the data into mutually exclusive or exhaustive groups.

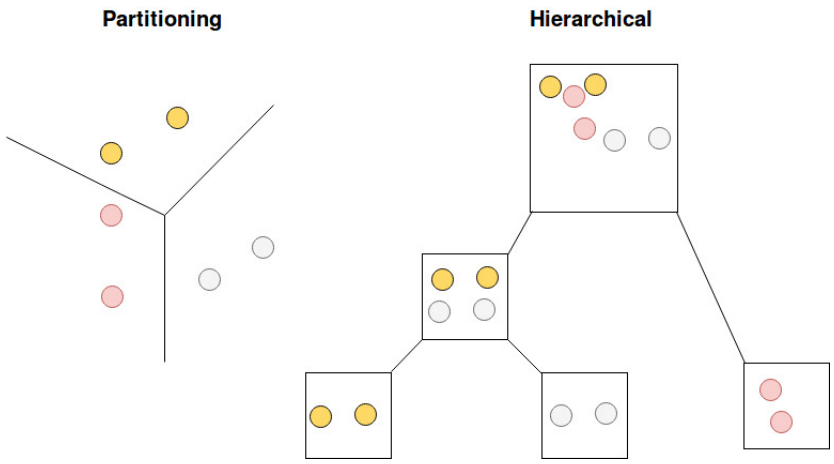


Figure 6: Two basics types of clustering

Clustering identifies the similar expression patterns under all conditions. This technique is used where each gene is defined by all conditions in the cluster or each condition is characterized by all of the genes. Traditional clustering methods like hierarchical clustering [23] and k-mean clustering [24] only group the rows based on their expression values in all conditions and one gene may have more than one function.

There are some possible drawbacks of the traditional clustering techniques. These are not effective when the genes perform differently across different experimental conditions. With such a scenario, it is more beneficial to use biclustering. This technique simultaneously clusters both rows and columns to identify the similarity of expression profile for determining the co-regulation of the genes. Biclustering selects each gene under the specific subset of the experimental condition where each condition is selected in the cluster for the specific number of genes. This method is used where only small set of genes are regulated in the cellular process of interest or specific cellular process, operated under some conditions. Some genes can be present in more than one cluster because genes can interact with different genes in different conditions.

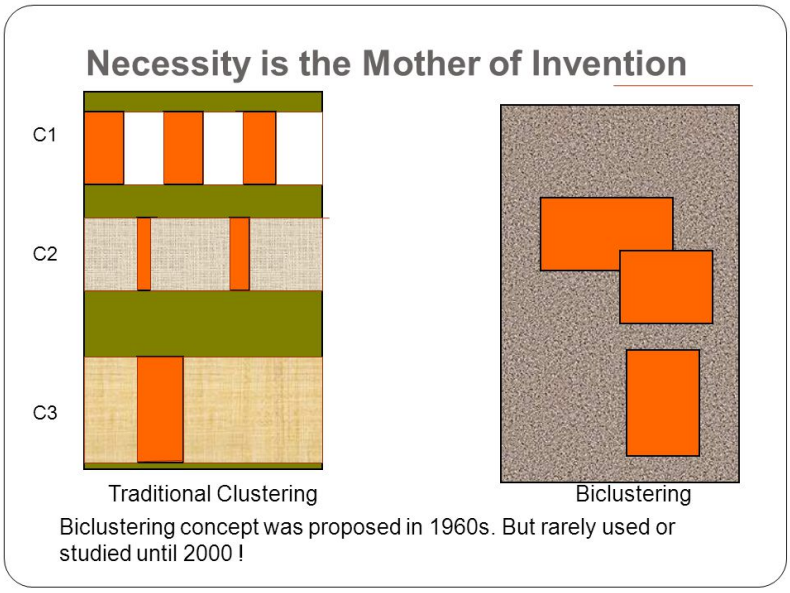


Figure 7: Clustering and Biclustering [4]

There are a number of proposed biclustering algorithms. Some of them are as follows:

- PLAID biclustering [25]
- Xmotif biclustering [26]
- Bimax biclustering [27]
- Spectral biclustering [28]
- Cheng and Church biclustering [29]
- FABIA Biclustering [30]
- SAMBA Biclustering [31]

FABIA

Factor Analysis for Bicluster Acquisition (FABIA) is based on factor analysis to find an unobserved variable (latent) by the observed variable [30]. This is a multiplicative model that assumes non-Gaussian signal distributions with heavy tails. FABIA implements model selection techniques like variational approaches and applies the Bayesian framework. The generative model helps the algorithm to distinguish between

a true cluster and spurious cluster. The multiplicative model considers two vectors similar if one is a multiple of the other. The overall model for p biclusters and additive noise is:

$$X = \sum_{i=1}^{p} \lambda_i \widetilde{z_i} + \epsilon = \Lambda \widetilde{z_i} + \epsilon \tag{16}$$

where ϵ is additive noise, p is the number of biclusters, λ_i is a sparse vector of factor loadings, and $\widetilde{z_i}$ is the i^{th} value in a vector of $\widetilde{z_i}$ factors. The Laplace Prior, Post-Projection, Sparseness Projection and SPARSE are available for FABIA. The laplace prior is mainly discussed in the following section and rest of them are described briefly.
FABIA with laplace prior have the products λ and Z which are Laplacian. Here we describe the distribution resulting from the product of two laplacian variables.

$$\boldsymbol{Y} = \Sigma_{p=1}^{P} \lambda_p Z_p + \epsilon \tag{17}$$

where Z_p is the *pth* factor, λ_p is the vector of factor loading for Z_p and additive random noise is assumed to be normally distributed, $\epsilon \sim N(0, \Psi)$. Furthermore, the model assumes that Ψ is a diagonal matrix and noise is independent from signal strength because of Z and Ψ are independent. Furthermore, the factor model assumes sparseness of factors and their loadings and this is reflected by the choice of the corresponding prior on loadings and factors (i.e. a Laplace Distribution). This factor model gives correct results on normalized data. The algorithm obtains clusters from λ_p Z_p and identification clusters are achieved by estimating the parameters and factors through λ, Ψ and Z. More intricate details and information about the method and its variations, can be found in Hochreiter et al. (2010) [30].

The Fabia Post-projection window is projected to a sparse vector according to Hoyer [32].

The prior of FABIA with sparseness projection is projected to finite support after each update of the loadings. This projection is done again according to Hoyer (2004) [32]. Again biclusters are discovered through sparse factor analysis and the model selection is performed by a variational approach according to Girolami (2001)[33] and Palmer et al. [34]. Furthermore a prior on the parameters is included and a lower bound on the posterior of the parameters is minimized, given the data. The update of the loadings includes an additive term which pushes the loadings towards zero (Gaussian prior leads to a multiplicate factor). More detailed information about this

algorithm and its methodology can be found in Hochreiter et al. [30].

PLAID Model

PLAID model is a flexible biclustering model which was introduced by Lazzeroni & Owen [25]. It describes the data matrix as a sum of K clusters and uniform noise whereas biclusters are called layers.

$$Y_{ij} = \mu_0 + \sum_{k=1}^{K} \theta_{ij}\rho_{ik}\kappa_{ik} \tag{18}$$

where μ_0 is background noise while $\theta_{ijk} = \mu_k + \alpha_{ik} + \beta_{jk}$ and u_k is the added background noise in biculster k. In bicluster k, α and β describe row and column specific effects. $\rho_{ik} \in 0, 1$ and $k \in 0, 1$ are indicator variables indicating the gene biculster membership and condition-bicluster membership of k^{th} bicluster. The overlapping may be prohibited by setting some constrains on κ and ρ. The aim of the model is to find the parameter values so that data could be fitted in the resulting matrix [35].
The above model can be described as a minimization problem.

$$argmax\left[\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}(a_{ij} - \theta_{ij0} - \sum_{k=1}^{K}\theta_{ijk}\rho_{ik}\kappa_{jk})^2\right] \tag{19}$$

Assume the $K-1$ clusters are found and want to find K clusters. The equation 19 can rewritten as following.

$$argmax\left[Q^{(K)} = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}](Z_{ij}^{(K-1)} - \theta_{ijk}\rho_{ik}\kappa_{jk})^2\right] \tag{20}$$

with

$$Z_{ij} = a_{ij} - \theta_{ij0} - \sum_{k=1}^{K=1}\theta_{ijk}\rho_{ik}\kappa_{jk} \tag{21}$$

To obtain the value of Z_{ij}, the values of θ, ρ and κ are updated iteratively in each iteration. To solve the equation, the algorithm proceeds as follows. The optimal value for θ^s is computed at each cycle given fixed values of $\rho^{(s-1)}$ and κ^{s-1} where iteration

S, ρ^0 and κ^0 is priory defined. Subsequently the optimal value for ρ^s is computed given $\theta^{(s)}$ and κ^{s-1} and value of $\kappa^{(s)}$ updated given ρ^{s-1} values. The value of κ, ρ and θ can be obtain by Lagrange Multipliers.

$$\mu_K = \frac{\sum_i \sum_j \rho_{iK} \kappa_{jK} Z_{ij}^{(K-1)}}{(\sum_i \rho_{iK}^2)(\sum_i \kappa_{jK}^2)} \tag{22}$$

$$\alpha_{iK} = \frac{\sum_j (Z_{ij}^{(K-1)} - \mu_K \rho_{iK} \kappa_{jK}) \kappa_{jK}}{\rho_{iK} \sum_{jK} \kappa_{jK}^2} \tag{23}$$

$$\beta_{iK} = \frac{\sum_i (Z_{ij}^{(K-1)} - \mu_K \rho_{iK} \kappa_{jK}) \rho_{iK}}{\kappa_{jK} \sum_{iK} \rho_{iK}^2} \tag{24}$$

For the minimization of equation 19, The optimal values of equation ρ and equation κ can be calculated :

$$\rho_{iK} = \frac{\sum_j \theta_{ijK} \kappa_{jK} Z_{ij}^{(K-1)}}{\sum_j \theta_{ijK}^2 \kappa_{jK}^2} \tag{25}$$

$$\kappa_{iK} = \frac{\sum_j \theta_{ijK} \rho_{jK} Z_{ij}^{(K-1)}}{\sum_j \theta_{ijK}^2 \rho_{jK}^2} \tag{26}$$

After S iterations, the cluster K would be rejected if its importance value is less than threshold. The cluster K would be accepted if its importance value is larger than the threshold. The importance can be calculated by $\sigma_K^2 = \sum_{i=1}^n \sum_{j=1}^m \rho_{iK} \kappa_{jk} \theta_{ijK}^2$.

2.2 ChIP-chip

One important protein in the columella stem cell niche is WOX5. WOX5 is a transcription regulator which binds a specific DNA sequence and regulates transcription at the binding site. Plants lacking WOX5 activity produce ectopic division of QC. Pi et al.(2015) [1] found that WOX5 represses CDF4 transcription in the stem cell niche. The CDF4 promoter is only known location where WOX5 binds to the DNA. ChIP-chip analysis is beneficial to find all additional potential binding sites of WOX5.

ChIP-chip or genome-wide location analysis [36] followed by genome tiling array employs chromatin immunoprecipitation (ChIP) is a powerful technique used for discovering the interaction between proteins and DNA via direct binding within the natural chromatin context of a cell. The aim of this technique is to determine the multiple protein interactions with a specific region of the genome like transcription factors on promoters or other binding sites. The ChIP-chip enables DNA microarray probes to tile the whole genome which generates one dimensional series of signals. The peak of those signals represent the protein binding sites. By identifying those peaks, the protein binding sites can be located which help to analyze the functional elements of the genome. The genome tiling array is different from a non-tiling array by composition. The tiled array selects a region of the genome and converts it into DNA fragments called tiled path. The average distance between each pair of neighbouring chunks (measured from the centre of each chunk) gives the resolution of the tiled path.

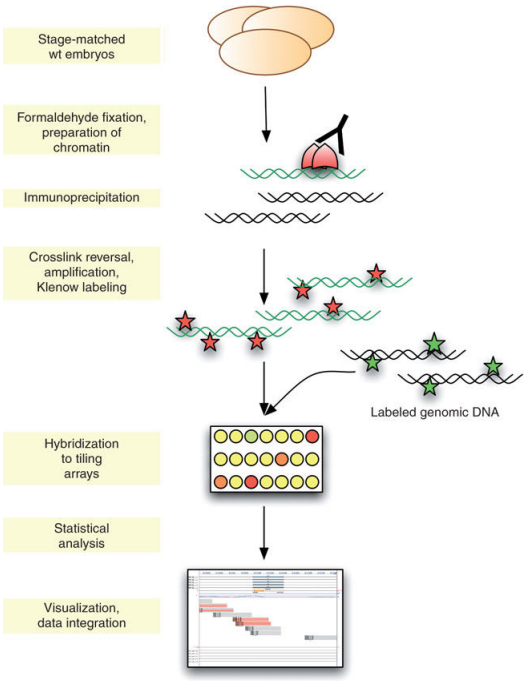
2.2.1 Workflow

The overview of the ChIP-on-chip workflow is presented in Figure **??**. In the first step, the protein of interest (POI) is cross-linked to DNA in vivo by chemical fixation methods. The cells are then lysed and the DNA is fragmented into small chunks, typically in 20-1000 base pairs by sonication. This gives output in double-stranded small pieces of fragments. The fragments that were cross-linked to the POI creates POI-DNA complex.

In the next step, the specific antibody is used against the POI-DNA complexes to remove them. The cross-links are reversed and the DNA fragments are purified. This procedure is called immunoprecipitation (IP) of the protein.

DNA is amplified by ligation-mediated polymerase (LM-PCR) chain reaction and labelled with fluorescent dye. The fragments are applied to DNA microarrays to detect the enriched signals. In the next step, Both IP-enriched and unenriched DNA

pools of labeled DNA are hybridized to the same chip. The workflow of hybridization and after hybridization is same as workflow of microarray described in section 2.1.



ChIP-Chip array Workflow [37]

2.2.2 Bayesian Modeling of ChIP-chip Data Through a High-Order Ising Model

Analyzing the ChIP-chip data is challenging because of the huge amount of probes, noise and its spatial dependency on probe intensity measurement. Bayesian Modeling of ChIP-chip Data Through a High-Order Ising Model [38] analyzes the ChIP-chip data by Bayesian hierarchical models approach in which the spatial dependency on probe intensity measurement models through ferromagnetic high-order or standard Ising models. This method identifies the transcription factor binding site for ChIP-chip experiment. The Bayesian Modeling of ChIP-chip Data naturally takes into account the intrinsic spatial structure of the data and can be used to analyze data from multiple platforms with different genomic resolutions [38].

Mo at el. [38] used hidden Markov Random Field-Based Bayesian model and Gibbs sampling for calculating the posterior probabilities to infer the binding sites of transcription factor. Briefly, without loss of generality, let each probe be associated with a binary latent variable $X_i \in 0, 1$ where i denotes the ID of the probe, X_i denotes the enriched probe and 0 otherwise. Let $\boldsymbol{y} = (y_1 \dots y_n)$ be a realization of enrichment measurements $\boldsymbol{Y} = (Y_1, \dots Y_n)$ along the chromosomes and each probe associate with a binary latent variable $X_i \in 0, 1$, where $X_i = 1$ denotes that the probe belong to binding region and 0 otherwise. We assume, conditional on X_i, Y_i follow the distribution

$$y_i|x_i \sim \begin{cases} N(\mu_0, \sigma_0^2), & \text{if } x = 0 \\ N(\mu_0, \sigma_1^2), & \text{if } x_i = 1 \end{cases} \tag{27}$$

In the first stage, conditioning on the latent variable, the probe enrichment measurements for each state (0 or 1) are modeled by normal distributions. In eq 27, Y_i is normally distributed with mean μ_0 and variance σ_0^2. The definition of Y_i is not too strict. The Y_i could be any appropriate measurement for the comparison of the IP-enriched and control samples. Here, this model doesn't account the signal specific effect of individual probe intensities rather it is designed to model the probe enrichment. Furthermore, it is assumed that conditional on $\boldsymbol{X} = (X_1, \dots X_n)$ $Y_i, ...Y_n$ are

independent. Thus we have

$$\pi(y|x,\psi) = \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi\sigma_0}} \exp\left(-\frac{(y_i-\mu_0)^2}{2\sigma_0^2} \right) \right)^{1-x_i} \tag{28}$$

$$\times \left(\frac{1}{\sqrt{2\pi\sigma_1}} \exp\left(-\frac{(y_i-\mu_1)^2}{2\sigma_1^2} \right) \right)^{x_i} \tag{29}$$

where $\boldsymbol{x} = (x_1 \dot{x}_n)$ is a realization of $\boldsymbol{X}$ and $\psi = (\mu_0, \sigma_0, \mu_1, \sigma_1)$ is a set of parameters.

The Priors

Let $\lambda_0 = \sigma_0^{-2}$ and $\lambda_1 = \sigma_1^{-2}$. To conduct a Baysian analysis, we assume that (μ_0, λ_0) and (μ_1, λ_1) are independent a prior, and they have the following prior densities:

$$\pi(\mu_0, \lambda_0) \propto \frac{1}{\lambda_0}, \pi(\mu_1, \lambda_1) \propto \frac{1}{\lambda_1} \tag{30}$$

The hidden state vector by a higher order Ising model is modeled as:

$$\pi(\boldsymbol{x}|\beta) = \frac{1}{Z(\beta)} \exp\left(\Sigma_{i=1}^{n} \left(\beta \Sigma_{j \in W(i)} \delta(x_i, x_j) \right) \right) \tag{31}$$

where $\beta > 0$ is the interaction parameter and $Z(\beta)$ is the normalizing constant of the distribution. If $\beta = 0$, then all probes are independent and there is no interaction between them. If $\beta < 0$, then model makes a spatial pattern of the neighboring probes in opposite state. If $\beta > 0$ is called ferromagnetic Ising model. $W(i)$ is a sliding window centered at probe i. This method probes information within a sliding window of certain gnomic distance. The $\delta(x_i, x_j) = 1$ if $x_i = x_j$ and -1 otherwise in model 31. The size of sliding window may be chosen on the probe resolution and if sliding window only consists of the immediately adjacent probes, the model 31 is reduced to the standard Ising model. Otherwise model turns into a higher-order Ising model.

Therefore for ChIP-chip it is necessary to restrict to positive value of β. This model 31 is called the ferromagnetic Ising model. We consider only the interactions among the nearest neighbors. When the interaction is beyond the nearest neighbors, the model turns into a higher-order Ising model.

The Full Conditionals and Gibbs sampling

In this section, the latent variable is modeled by ferromagnetic Ising model. The posterior probability of model parameters is used for simulation of the Gibbs sampler and metropolis algorithm. The probe with high posterior probability of a enriched state means that it is strong evidence that the probe belong to binding region.

Let $n_0 = \Sigma_{i=1}^{n}(1 - x_i)$ and $n_1 = \Sigma_{i=1}^{n}(x_i)$ be the total numbers of non-enriched and enriched probes, respectively. Let $\bar{y_0} = \frac{1}{n_0}\Sigma_{i=i}^{n}(1 - x_i)y_i$ and
$\bar{y_1} = \frac{1}{n_1}\Sigma_{i=i}^{n}x_i y_i$ be the sample means of the measurements for non-enriched and enriched probes, respectively. In addition, let $N(a, b)$ denote a Gaussian distribution with mean a and variance b, let $Ga(a, b)$ denote a gamma distribution with mean a/b and variance a/b^2, and let $y \mid ..$ denote the full conditional distribution of y given everything else in the model. After some algebra, we get the following full conditional distributions:

$$\mu_0|\cdots \sim N(\bar{y_0}\frac{1}{n_0\lambda_0}) \tag{32}$$

$$\lambda_0|\cdots \sim Ga\left(\frac{n_0}{2}, \frac{1}{2}\Sigma_1^n(1 - x_i)(y_i - \mu_0)^2\right) \tag{33}$$

$$\mu_1|\cdots \sim N(\bar{y_1}\frac{1}{n_1\lambda_1}) \tag{34}$$

$$\lambda_1|\cdots \sim Ga\left(\frac{n_1}{2}, \frac{1}{2}\Sigma_1^n(1 - x_i)(y_i - \mu_1)^2\right) \tag{35}$$

$$\pi(x_i = 1|\dots) = \left(1 + \left(\frac{\lambda_0}{\lambda_1}\right)^{\frac{1}{2}} \exp\left(\beta(n_0(i)) - n_1(i)) + \frac{\lambda_1}{2}(y_i - \mu_1)^2 - \frac{\lambda_0}{2}(y_i - \mu_0)^2\right)\right)^{-1}, \tag{36}$$

where $i = 1\dot{n}$, and $n_o(i) = \Sigma_{j\in W(i), j\neq i}\delta(0, x_j)$ and $n_1(i) = \Sigma_{j\in W(i), j\neq i}\delta(1, x_j)$ are the sums of interactions given probe i is a nonbinding or a binding probe in the sliding

window $W(i)$, respectively. Given the full conditional distributions eq 32 and eq 36, it is straightforward to simulate from the posterior distributions using a cyclic Gibbs sampler. The posterior probabilities of the hidden states will be used for inference of binding sites. Because the model is not fully identifiable, the likelihood of staying the same by switching (μ_0, λ_0) with (μ_1, λ_1) and switching all of the cluster labels (0 by 1 and 1 by 0). The outputs have been relabeled according to the constraint $\mu_0 < \mu_1$.

More detailed information about this algorithm and its methodology can be found in Mo et al. [38].

3 Related Work

The transcription network specifies the cell types in animals and plants. It is important to map the transcriptional network for the deep understanding of the spatial and temporal control of organs. Brady et al. [39] profiled nearly all cell types of Arabidopsis thaliana by using the microarray data and used a set of co-expressed genes to map the transcriptional network.

Only small information is available about the molecular mechanisms which determines the functionalities of the QC or initial stem cells. Nawy et al. [40] analyzed the gene profile of QC and provided the information about the genes which maintain and contribute to the activity of QC cell. The authors analyzed microarray data by linear mixed-model analysis of variance to identify the set of QC enriched genes and showed that it is possible to profile a cell type.

WOX5 is an important protein in the columella stem cell niche which acts as a stem cell regulator and maintains QC undifferentiated. Pi et al. [1] showed that WOX5 protein moves from the root niche organizer to columella stem cells with the higher level in the QC, a weaker level in CSC, and no regulation in CRC and controls QC division. CDF4 gene is a direct target of WOX5 in these cell types which form a reverse gradient of WOX5 in stem cell niche.

Another notable technique which can be used in transcriptome profiling is biclustering. Asyali et al. [41] discussed the biclustering as a technique to identify the subset of genes which may participate in any cellular process under a variety of different experimental conditions or gene may be involved in multiple biological processes. The authors described the importance of clustering in both genes and conditions simultaneously which could help in identifying co-regulated genes, protein interactions and transcription factor binding site to identify modules.

For identification of the gene clusters which could help the understanding of genes common regulatory networks, Kluger et al. [42] proposed a spectral biclustering method. The main underlying assumption of this method was that it is possible to cluster a subset of overrepressed marker genes of one tumour which are not overexpressed in other tumours.

4 Approach

In this section, the implementation details of the methods are described which were mentioned in the section 2.

4.1 Data insight

The four replicates of quiescent center(QC) cells were read-out from two-channel microarray labelled with pWOX5::GFP. We are only interested in the wild type of cell types for the transcription profiling. Therefore, only wild-type from the two-channel microarray was processed for further analysis.

For single-channel array, the samples were dyed only with green fluorescent. We analyzed the sample four replicates of QC, four replicates of CSC and eight replicates of CRC. from single-channel array.

For two-channel array, the expression values from red intensities were omitted from aglient raw data. The expressions from green intensities of single channels and two-channel microarray are combined for bioinformatics analysis.

In the aglient raw data, the respective genes assigned to different values (-1,0,1) on microarrays are called control type. The control type value -1 designates the probe group as a negative probe group which are intended to have no hybridization. The control value 1 designates as positive user control and indicates generally predictable signals. These signals are used for downstream analysis but most of the times these are excluded from statistical analysis. The default value is 0 which means that the probe group is not of any control type. This default value was chosen for control type in our analysis because only hybridized signals are meaningful for further procedure.

Marker genes

As a further test of the cell-specific expression, our former lab member, Dr. Ernst Aichinger, cloned the promoters of a set of genes predicted to be enriched in the QC, CSC and CRC (Aichinger, unpublished) [43]. He fused them to GFP and introduced

the constructs into plants. GFP exhibited expression in the QC as well as CSC and CRC.

Along with Dr. Aichinger marker genes, we also used some cell-specific genes of QC, CSC and CRC from previous literature to support the validity of the columella stem cell niche transcriptional profile. These genes are given in Table 17, Table 18, Table 19 and Table 20 in supplement. For simplicity, we shall refer all of these genes as Dr. Aichinger marker genes.

4.1.1 preprocessing

There are many reasons for the variation in the intensity of the microarray. Some of the reasons are different scanner settings, the difference in room temperatures and technician's experience level. These variations are considered as a bias. The other type of bias is probe effect, such that probes outside the binding region have higher intensities compared to the inside bounding region.

There are a number of steps for removing biases of microarray data. The first step is background correction.

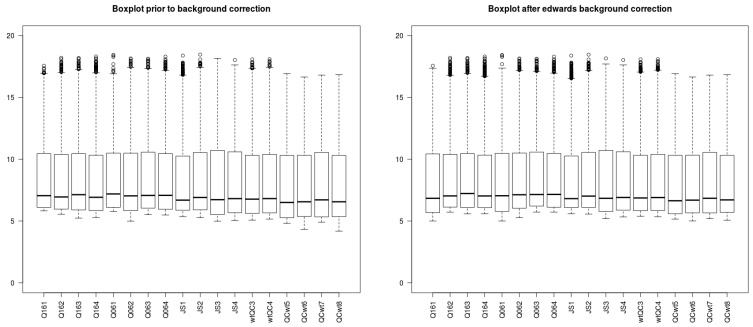
Background correction

The background correction is necessary for removing the effects of non-specific binding. There are a number of methods which are available for background correction. Four commonly used methods are 'normexp' [13], 'subtract', 'edwards' [14] and 'movingmin'. We applied these methods for background correction and these methods are available in R package limma [44] for microarray data. Normexp [13] is the first method which we considered for the background correction. This method is recommended for two channel arrays, since our data only consists of single-channel array intensities, this method was not suitable for our case.

Other methods are 'subtract', 'edwards' and 'movingmin'. These methods were applied on the data set for background correction and all of these methods produced similar outputs (as shown in Figure 8).

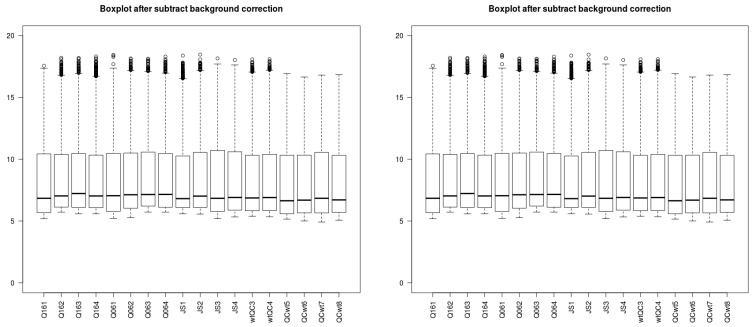
We corrected the background by edwards method for further data processing because in our experience, edwards cause few problems after background correction of normalization of data. As edwards is a simple approach for adjusting the background, it applies the smooth monotonic function for adjustment if the difference between foreground intensities and background intensities is less than a threshold [14]. Con-

trarily if the difference is larger than the small threshold then it simply subtracts the background according to foreground [18].



(a) Intensities after prior to background correction correction

(b) Intensities after edwards background correction



(c) Intensities after movingmin background correction

(d) Intensities after subtract background correction

Figure 8: Boxplots displaying the intensity distribution for different replicates before background correction and after edwards method, movingmin method and subtract method background correction. The first 6 boxplots represent the data for the CRC cell and the next 4 represent the data for the CSC and last 6 boxplots represent the data for QC.

Normalization

Normalization is an essential procedure in microarray data analysis. Normalization is a process to identify and remove the systematic source of variation in a microarray experiment which affects the measured gene expression levels. The reason for the normalization is to form same distribution for each array to make multiple arrays comparable. As a first step, one needs to decide which method is to use for data normalization. A number of different normalization methods are developed and

tested for spotted microarrays [15][16]. Symth and Speed [17] describe some of most commonly used methods. These methods remove the technical systematic biases of log ratios from one array or between arrays. The reason for the existence of different methods is because there is no gold method which can correct all technical artifacts. It is relatively difficult to select one method for normalization because there is the non-existence of a total signal in real biological data. These normalization methods could be classified into two methods.

- Within-Array Normalization.
- Between-Array Normalization.

Within-Array Normalization normalizes the M values for two-color spotted microarray. It is usually used for traditional log ratios of two-color data. For single-channel arrays, within-array normalization is not usually relevant.

Between-Array normalization is the sole normalization step for single-channel microarray. In our case, Between-Array normalization is the better option because our data only consists of one-color data and from multiple arrays. For single-channel data, quantile, scale and cyclicloess normalization methods work better because when an object is a matrix or EListRaw object from single-channel arrays, other methods produce an error. The quantile and cyclicloess normalization was originally proposed by Bolstad et al. [18] for Affymetrix-style single-channel arrays. These three methods are available in R package limma [44], The quantile and cyclicloess are called complete methods because these methods make use of data from all arrays in an experiment to form the normalizing relation.

The quantile, scale and cyclicloess were applied for normalizing the expression values after correction of background. The "scale" method is the simplest method for normalization. This method simply scales the columns to have the same median (set the median of difference to 0). The media centering methods make the assumption that the majority of genes are unchanged between the conditions.

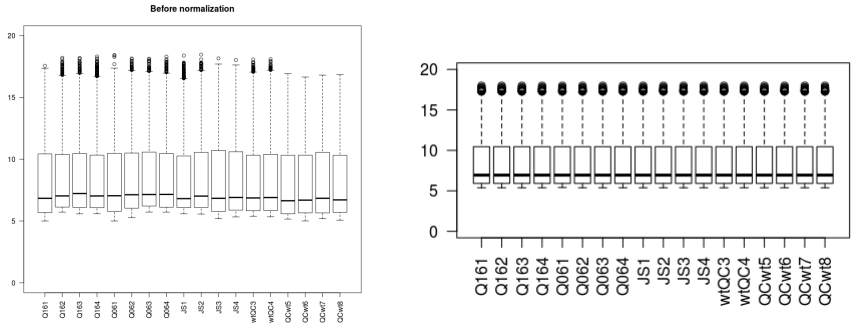
The "quantile" method tightens the idea of "scale" method. It ensures that the intensities have the same empirical distribution across arrays and across channels [45].

It is difficult to show the MA plots of every replicate for each normalization method because of limitation of page space. We selected only one representative replicate for each of the normalization methods. The MA plot was plotted for representative replicate by averaging all the arrays other than representative replicate.

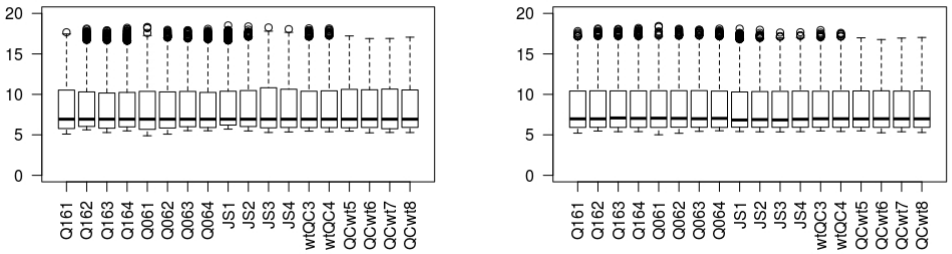
The normalized intensities by quantile and scale methods were not chosen for further analysis. As it is shown in Figure 9(b) and Figure10(c) that quantile method over normalizes the values and forces the values of quantiles to be equal and not only adjusted to 50% quantile but all quantiles. This could be most problematic in the tails where it is possible that a probe could have the same value across all the arrays.

The scaled normalized data is not chosen for further analysis because as Figure 9(c) and Figure10(d) shows it under normalization the expression values and Figure 11(c) shows that MA plot of scale normalization shows a little oscillation.

We normalized our background corrected data by cyclicloess. Cyclicloess is a loess based method and implements the idea of log-based ratios, presented in Dudiot et al. [46]. Basically, cylicloess use log ratios M and A for analyzing the expression data where M is the difference in log expression values and A is the average of log expression values. This method systematically cycles through all the pairs of probe many times and probe intensities of two arrays rather than the two-color array.
The advantage of this normalization method is that it is well suited for tackling the unbalanced differential expression and as Figure 9(d) and Figure10(b) shows that it didn't under or over normalize the expression values.

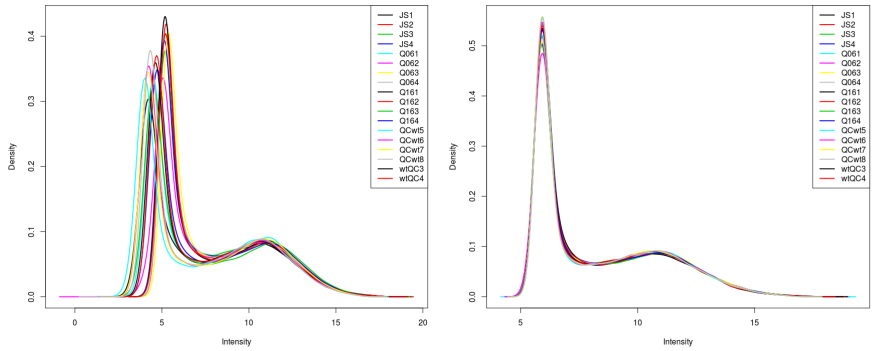


(a) Intensities prior to normalization **(b)** Intensities after quantile normalization

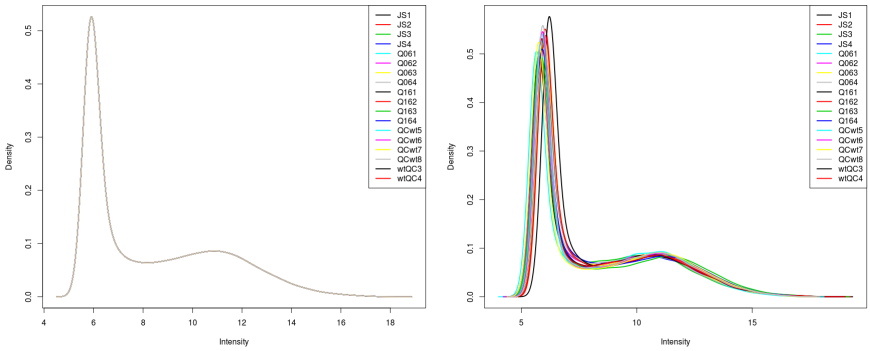


(c) Intensities after scale normalization **(d)** Intensities after cyclicloess Normalization

Figure 9: Boxplots displaying the intensity log ratio distribution for different replicates after quantile, scale and cyclic loess normalization. The first 6 boxplots represents data for the CRC cell and the next 4 represent data for the CSC and last 6 boxplots represent data for QC.



(a) Intensities before normalzation **(b)** Intensities after cyclicloess normalization



(c) Intensities after quantile normalization **(d)** Intensities after scale normalization

Figure 10: Density plots replicates densities. Color curves represent the densities of different replicates

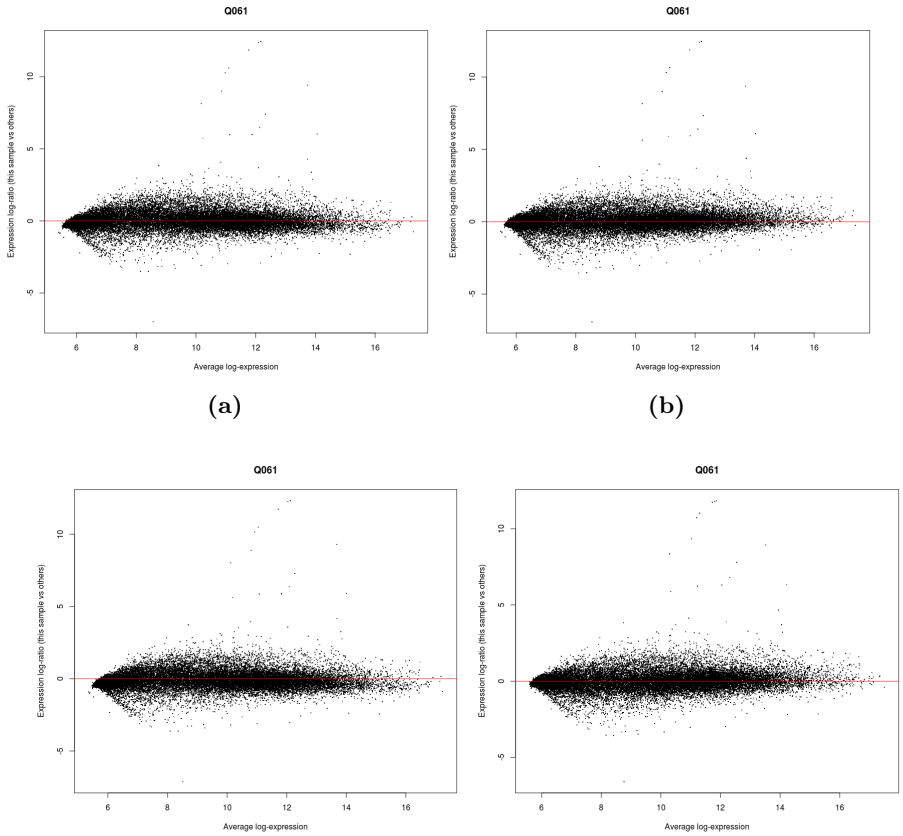


Figure 11: Post-normalization MA plots. MA plots show log fold changes (M) as a function of the mean single channel intensity (A). The horizontal red line shows the median of the M-values.(a) MA plot of a representative sample prior to normalization. (b)MA plot of a representative sample after cyclicloess normalization. (c) MA plot of a representative sample after scale normalization. (d) MA plot of a representative sample after quantile normalization.

4.2 Differential Expression Analysis

Linear models are among the most used statistical methods for modelling and data analysis. The t-statistics, ANOVA and regression are special cases of the linear models. The linear models define a linear relationship between observed values and experimental conditions. These model can summarize the log-ratio of each gene for designed microarray experiment with some level of replication before testing

for differential expression [19] and compare two groups or multifactorial designs e.g genotype or treatment.

After our data was normalized and available as log ratios of each gene for each condition, the linear model was fitted on expression data for comparing RNA sources of these QC, CSC and CRC.

We used R package limma [47] for implementation of the linear model. This model fits the linear model to expression data for each probe and accepts the log ratio as a data set. This method is the implementation of Symth et al. [19]. The authors reset the hierarchical model of Lonnstedt and Speed [21] in the context of general linear models for identification of differentially expressed genes. This method needs design matrix and contrast matrix to be described. The design matrix has a description of hybridized RNA and contrast matrix uses the coefficients of design matrix for comparison of differences between RNA sources hybridized to the array. The coefficients of the fitted models depict the contrasts between the RNA sources. For single-channel array data, the design matrix can be designed just like modelling univariant data and the linear modelling can act like ordinary univariate linear models and ANOVA except model is fitted for every gene.

The aim of the cell types comparison analysis is to identify the enriched genes in the columella stem cell niche. For comparison, we made certain contrast matrix where contrast was compared one-by-one with each other. After the model was fitted to the data, the empirical Bayes was calculated.

After empirical Bayes processing the result of the linear model, the list of deferentially expressed genes were extracted. The p-value was adjusted for multiple testing by FDR method where FDR controls the expected proportion of false discoveries under the specified values [48].

To determine enrichment, we applied adjusted p-value below 0.05 value and log fold above 1 value for every pairwise comparison. The genes below the 0.05 value are considered statistically significant genes. The statistically significant DEGs were obtained for stem cell niche cell types. The list of comparisons is following where vs indicates the comparison between two cell types.

- QC vs CSC
- QC vs CRC
- CSC vs QC
- CSC vs CRC

- CRC vs QC
- CRC vs CSC

By adjusted p-value and log fold change, genes of QC vs CSC and QC vs CRC were combined to find the QC specific genes, genes of CSC vs QC and CSC vs CRC were combined to find the CSC specific genes and genes of CRC vs CSC and CRC vs CSC were combined to find the CRC specific genes.

Furthermore, we identified the increasing gradient and decreasing gradient in stem cell niche. The decreasing gradient of a gene could be defined as a gene which has higher log fold change value in QC compared to CSC and higher log fold change value in CSC compared to the CRC. Similarly, increasing gradient of a gene which is defined as a gene has lower log fold change value in QC compared to CSC and lower log fold change value CSC compared to CRC.

We extracted statistically significant genes of CSC vs CRC with log fold change above 1 and QC vs CSC with log fold value above -1 and combined them by log fold change to obtain the increasing gradient in stem cell niche cell types. Similarly statistically significant genes of QC vs CSC with log fold change value below -1 and CSC vs CRC with log fold change value below 1 log fold change were combined to obtain increasing gradient. This gradient was further combined with the WOX5 binding list.

4.2.1 Biclustering

Biclustering is an important technique to identify the function related genes under different conditions. Basically, this technique finds the clusters or groups of genes with the similar pattern of expression values under different conditions and those genes can share more than one cluster. This approach is appealing in biology, a number of methods are developed for the biclustering. We evaluated following well-established algorithms to find the cell-specific genes of QC, CSC and CRC.

- PLAID biclustering [25]
- xMOTIFs biclustering [26]
- Bimax biclustering [27]
- Spectral biclustering [28]
- Cheng and Church biclustering [29]

- Factor Analysis for Bicluster Acquisition [30]
- Qualitative Biclustering [49]

First candidate algorithm for our data was Cheng and Church biclustering [29] which minimizes the mean squared residue by adding or removing the rows and columns.

Another candidate algorithm was spectral clustering [28]. It uses the single decomposition value on normalized microarray data to find the checkboard pattern.

The Binary Inclusion-Maximal Biclustering Algorithm (Bimax) [27] uses the divide and conquer approach for searching the submatrices formed with entries whose values are equal to one.

The Qubic algorithm (QUalitative BIClustering algorithm) [49] is a biclustering algorithm which clusters the data by qualitative manner. It uses statistical models to find all statistically significant biclusters .

Conserved Gene Expression Motifs (xMOTIFs) [26] is a non-deterministic algorithm that finds submatrices with simultaneously conserved genes in subsets of experimental conditions based on the greedy approach.

We applied all 7 biclustering algorithms which are mentioned above on our transcriptome dataset. The validation criteria for the algorithm were Dr. Aichinger marker genes [43] for QC, CSC and CRC specific clusters. The first algorithm which was applied to the dataset was Cheng and Church algorithm. This algorithm failed to find biclusters with any overlap because of its masking technique. The algorithm was not selected for further analysis because generally overlapping biclusters represent true nature of the genes. The xMOTIFs, Bimax, spectral clustering and QUBIC were applied with their different settings of parameters but none of them was validated by marker genes. Only the two following algorithms were validated by Dr. Aichinger marker genes [43]

- PLAID
- FABIA

PLAID

The PLAID algorithm was applied to the transcriptome dataset consists of the union of the cell-specific DEGs of a stem cell niche cell types obtained from differential expression analysis. This algorithm was validated by Dr.Aichinger marker genes [43]. The number of Dr. Aichinger marker genes [43] for QC, CSC and CRC were 14, 9

and 16 respectively. The R package Biclust [50] was used for implementation of this algorithm. The PLAID algorithm in Biclust package was proposed by Turner et al [51] which is improved version of the original PLAID algorithm [25]. The algorithm was applied to two datasets. The first dataset has cell-specific DEGs of stem cell niche cell types with the adjusted p-value below the 0.05 and second dataset with the adjusted p-value of same cell types below the 0.01. Each DEG was assigned to its normalized expression values under all conditions.

The goal of the algorithm implementation was to find QC cell-specific cluster, CSC cell-specific cluster and CRC cell-specific cluster which completely contained QC, CSC and CRC Dr. Aichinger marker genes [43] respectively.

Table 1 gives a short description of parameters of PLAID. After running the PLAID algorithm, it returns an object of the class which contains some values for obtaining the biclusters. The overview of the returned object is shown in Table 2.

There is no such thing like optimal parameters setting for all types of dataset. However, it is possible to find appropriate biclusters by giving the correct parameter setting.

For finding the correct parameter setting for identifying the desired biclusters in our dataset with the adjusted p-value < 0.05, parameters of PLAID were tuned by random search. The termination criteria for the algorithm was to find more than three clusters in which one cluster contains all of Dr. Aichinger QC maker genes, one cluster which contains all Dr. Aichinger CSC maker genes and one cluster contain Dr. Aichinger CRC maker genes. At first, two parameters `iter.startup` and `iter.layer` were initialized with the random position in search space keeping rest of parameters at the default position. `iter.startup` is the number of iterations and `iter.layer` is a the number of iterations to find each layer/cluster. There is no upper bound or lower bound for these parameters. During the random search, it was seen that the results get improved when the values of `iter.startup` are 1 and 64. 1 was assigned to the value of `iter.startup` and initialized random search again only for the parameter `iter.layer`. The termination criteria met with the `iter.layer` value of 94.

After that, we again applied PLAID on the data set with adjusted p-values < 0.01. For tuning the parameters in this data set. We assigned `iter.startup` from 1 to 1000 values randomly. We found results according to our termination criteria when `iter.layer` was 1 and `iter.startup` was 239. The clusters were further subjected to Gene ontology analysis.

Parameters	Details
`x`	The data matrix in which biclusters should be found.
`method`	Choice of the method. The argument `BCPlaid()` performs the PLAID algorithm.
`cluster`	`"r"` clusters rows,`"c"` columns and `"b"` (default) both.
`fit.model`	Linear model to be fit. The formulae is similar to the θ parameter in the model description. `m` is the overall bicluster constant μ,a the row constant α and b the column constant α.
`background`	If `TRUE` the function will allow an overall background layer in the data matrix.
`row.release`	Threshold to prune rows in the layers. Scalar in `[0,1]` with recommended interval `[0.5,0.7]`
`col.release`	Same as row.release, but for columns.
`shuffle`	Number of random layers to compute the significance of an observed layer. Default is set to 3.
`back.fit`	Additional iterations to refine the fitting, after a layer was found (default set to 0).
`max.layers`	Maximum number of bicluster to be found.
`iter.startup`	Number of iterations to find starting values.
`iter.layer`	Number of iterations to find a bicluster.
`verbose`	If `TRUE` extra information on progress is printed.

Table 1: Overview of the parameters of `"biclust"` function

Slot	Details
`Parameters`	Contains a list of the input parameters.
`RowXNumber`	Logical Matrix which gives the row bicluster membership.`TRUE` in `[i,j]` if row `i` is in bicluster `j` .
`NumberXCol`	Same as `RowXNumber`, but for columns. `TRUE` in `[ i,j ]` if column `j` is in bicluster `i` .
`cluster`	`"r"` clusters rows, `"c"` columns and `"b"` (default) both.
`Number`	Number of observed bicluster.
`info`	Additional information on bicluster. For example Sum of Squares(SS) and Mean Sum of Squares(MS).

Table 2: Overview of the return values of class `"biclust"` function

FABIA

FABIA [43] was second biclustering algorithm which generated the clusters validated by Dr. Aichinger marker genes. The FABIA is a model-based technique which uses multiplicative model and generative framework to determine the biclusters.

Although marker genes validated the clusters generated by PLAID, there was still some scope to improve the clusters. Therefore FABIA was applied. It was the second biclustering algorithm which identified apparently true clusters for stem cell niche. FABIA was applied to same data sets as PLAID in section 4.2.1. The R package "FABIA" [30] was used. The default parameters of FABIA package are following and description of parameters is given in Table 3:

```
fabia(X,p=5,alpha=0.1,cyc=500,spl=0,spz=0.5,random=1.0,
      center=2,norm=1,scale=0.0,lap=1.0,nL=0,lL=0,bL=0)
```

We tuned the parameters of FABIA by the random approach. First, we assigned `center=1` because we wanted to apply mean method for the data centering. We initialized the algorithm with `p=10` because termination criteria for PLAID was met with 10 clusters and afterwards we varied the `p` value between 6 to 18. The termination criteria was the same as we defined for the PLAID in section 4.2.1. We found true clusters by following parameters of FABIA for the dataset with adjusted

p-value 0.01 cutoff.

```
fabia(X,p=17,alpha=0.1,cyc=2314,center=1)
```

We applied the same approach for tuning the parameter for data set with adjusted p-value 0.05 cutoff as we applied for the dataset with the adjusted p-value of 0.01. Our clusters were validated by marker genes with the following parameters.

```
fabia(X,p=17,alpha=0.1,cyc=3487,center=1)
```

The clusters are further analyzed by GO analysis.

Parameters	Details
`X`	The data matrix.
`p`	number of hidden factors = number of biclusters; default = 5.
`alpha`	sparseness loadings (0 - 1.0); default = 0.1.
`cyc`	number of iterations; default = 500.
`spi`	sparseness prior loadings (0 - 2.0); default = 0 (Laplace)
`spz`	sparseness factors (0.5 - 2.0); default = 0.5 (Laplace) with recommended interval `[0.5,0.7]`
`random`	random initialization of loadings in [-random,random]; default = 1.0.
`<=0:by SVD,`	
`>0:`	
`center`	data centering: 1 (mean), 2 (median), > 2 (mode), 0 (no); default = 2.
`norm`	data normalization: 1 (0.75-0.25 quantile), >1 (var=1), 0 (no); default = 1.
`scale`	loading vectors are scaled in each iteration to the given variance. 0.0 indicates non scaling; default = 0.0.
`lap`	minimal value of the variational parameter, default = 1.
`nL`	maximal number of biclusters at which a row element can participate; default = 0 (no limit)
`lL`	maximal number of row elements per bicluster; default = 0 (no limit).
`bL:`	cycle at which the nL or lL maximum starts; default = 0 (start at the beginning).

Table 3: Overview of the parameters of the `FABIA` function

Return values of the algorithm:

Object of the class `Factorization`. Containing `LZ` (estimated noise free data $\boldsymbol{\Lambda Z}$), `L` (loadings $\boldsymbol{\Lambda}$), `U`(noise $\boldsymbol{X}-$), `center` (centering vector), `Z` (factor $\boldsymbol{Z}$), `Psi` (noise variation Ψ), `lapla` (variational parameter), `avini` (the information which the factor z_{ij} contains about averaged over j), `xavini` (the information which the factor $\widetilde{z_j}$ contains about averaged over j), `ini` (for each j the information which factor z_{ij} contains about $\boldsymbol{x_j}$).

4.2.2 Gene Ontology Analysis

The biclusters obtained from the PLAID and FABIA methods were further subjected to Gene Ontology (GO) analysis. In order to gain insight into the developmental processes operating in the three cell types of the columella stem cell niche, only the biological process (BP) domain of GO was considered. The inputs for GO analysis are simply the systematic names of the genes found in each bicluster. For both biclustering methods, more biclusters than cell types were obtained; therefore, only the biclusters that best represented each cell type were used. Each biclustering method was run on genes that differed at the 0.05 and 0.01 significance levels among the cell types, resulting in four sets of GO analyses on cell types of the columella stem cell niche.

Many packages and website tools can perform GO analysis, for example DAVID [52, 53], GORILLA [54] and PANTHER [55]. These three Web-site tools support Arabidopsis, but DAVID was chosen for several reasons. The DAVID database is regularly updated. The standard AGI (Arabidopsis Genome Initiative) systematic names are accepted as input, and the full transcriptome is available as a background data set. Not only GO terms can be obtained from DAVID, but also enrichment in gene annotation, pathways, protein domains, protein interactions and key words can be added to the analysis. Finally, the resulting tables can be downloaded in tab-delimited-text format, facilitating downstream interpretation and analysis in spreadsheets and in the R statistical software.

Each bicluster of gene systematic names was uploaded to DAVID, and only the GOTERM_BP_DIRECT annotation was selected for enrichment analysis. The 'Funtional Annotation Chart' view was chosen, and the options changed to include fold enrichment in the chart. The table was downloaded as a tab-delimited-text file and imported into R. The GO terms (Term) and adjusted p-values (Benjamini) of

the terms significant at the 0.05 level were saved in another tab-delimited-text file and uploaded to REVIGO [56, 57]. The REVIGO tools uses semantic similarity to collapse related GO terms into a representative GO term. The REVIGO options were set to: small resulting list, p-values, Arabidopsis thaliana GO database, and the SimRel (Schlicker's Relevance) semantic similarity measure. In the 'Scatterplot Table' tab, the biological process table was saved as a csv file and only the representative terms (eliminated=0) were further interpreted.

4.3 ChIP-Chip Analysis

4.3.1 Data Insight

Chromatin immunoprecipitation followed by tiling microarray analysis can identify potential DNA binding sites of a transcription factor. Our aim of this experiment was to identify the binding sites of WUSCHEL-RELATED HOMEOBOX 5 (WOX5) in the stem cell niche.

The experiment was done with two replicates under two conditions (input where no antibody replaces control with input without any IP enrichment and IP-enriched which use WOX5 antibody) for measurements. The two replicates are the minimum number of replicates which is recommended for ChIP-chip analysis. The reasons for this recommendation is to detect the corrupted measurements which are easy to identify in multiple replicates. Another reason is unsystematic noise of standard deviation can be minimized by averaging over samples. In the experiment, the transcription factor WOX5 was used for WOX5 specific IP. Following sample hybridization, the arrays were scanned by laser to scan the fluorescence and the fluorescence intensities were converted into an image which was used in the bioinformatics analysis.

4.3.2 Data Prepossessing

We went through the procedure of quality control for the data similarly as we did in section 4.1 and we didn't find any problem in the data for correction. The "normexp" and "cyclicloess" methods were selected for background and normalization correction respectively.

Data Analysis

The R package "iChip" was used for analyzing the ChIP-chip data [58]. There were two samples for analysis. The ChIP-chip data was in log ratio format after background correction and normalization.

At first, the IP-enriched and control samples were compared for enriched measurement using empirical bayes t- statistics method to process data further for detecting the enriched region. The enriched region was detected by second-order Ising model using the posterior probability. There is a function in the "iChip" package for applying second-order Ising model. The function used in this work can be called with the

following command.

```
iChip2(Y=oct4Y,burnin=2000,sampling=10000,winsize=2,
                    sdcut=2,beta=1.25,verbose=FALSE)]
```

The following are parameters for the iChIP function.

Slot	Details
`Y`	A n by 2 matrix or data frame
`burnin`	The number of MCMC sampling iterations. The posterior probability of binding and non-binding state is calculated based on the samples generated in the sampling period.
`sampling`	The number of MCMC sampling iterations. The posterior probability of binding and non-binding state is calculated based on the samples generated in the sampling period.
`winsize`	The parameter to control the order of interactions between probes. For example, winsize = 2, means that probe i interacts with probes $i-2, i-1, i+1$ and $i+2$.
`sdcut`	A value used to set the initial state for each probe. The enrichment measurements of a enriched probe is typically several standard deviations higher than the global mean enrichment measurements.
`beta`	The parameter used to control the strength of interaction between probes, which must be a positive value.
`verbose`	A logical variable. If TRUE, the number of completed MCMC iterations is reported.

Table 4: Overview of the return values of the class "`biclust`" function

We applied the Higher Ising model to the normalized dataset. The normalized dataset consisted of log2 ratios of the intensities of IP-enriched and control samples of probe enrichment measurement for a single replica, First column of data contained the chromosome IDs; the second column contained the probe enrichment measurements.

Data was sorted, first by chromosome and then by genomic position. After that, The parameters of the function were tuned for finding the optimal parameter settings. The value for `winsize=2` was recommended to balance between high sensitivity and to obtain low FDR. For checking whether the model parameters converge or not, the model parameters were plotted. It was recommended in the "iChIP" documentation

that if the parameter is continuously increasing or decreasing then the parameters are further required to be tuned. It was seen that by increasing the iterations in the burn-in phase. The plotted model parameters kept improving until the `20000` iteration. After 20000 iterations, plotted model parameters stopped improving. We fixed the `burnin`=20000. After that we started tuning the parameter `beta`. The increasing value of beta means less enriched regions and more strict criteria for detecting the enriched region. The required plotted model parameters were achieved by fixing the value of `beta` 1.7. The plotted parameter model is shown in Figure 13 for tuned parameter settings.

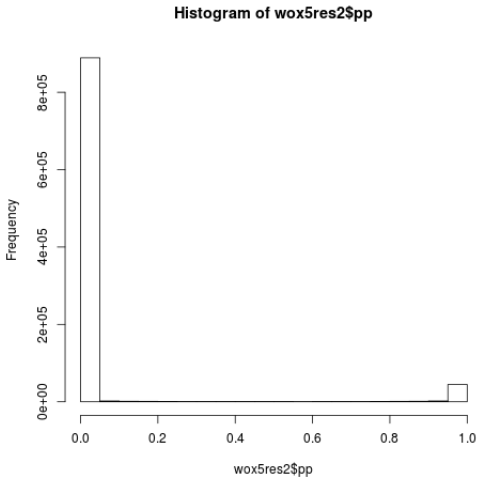


Figure 12: Histogram of the posterior probability of enriched binding region for WOX5. The WOX5 binding sites are dominated by 0 and WOX5 binding sites has 1 posterior probability

The WOX5 enriched region was called by posterior probability from the output of Higher Ising model. The detected enriched region of WOX5 was cut off by two types of methods, FDR [59] and posterior probability. In case of transcription factor binding region, the posterior probability should be dominated by 0 values. The histogram of our results is shown in the Figure 12

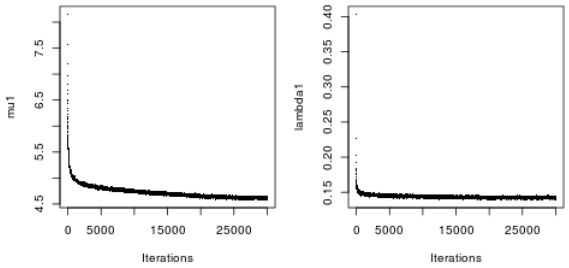


Figure 13: Plots of the parameters trend

The results were obtained of the enriched region by FDR cutoff of 0.01 and the posterior probability of 0.9. The results are shown in the result section 5

`chr`	Chromosome IDs.
`gstart`	The start genomic position of the enriched region.
`gend`	The end genomic position of the enriched region.
`rstart`	The row number for gstart in the position matrix.
`rend`	The row number for gend in the position matrix
`peakpos`	The peak genomic position of the enriched region where the probe has the largest enrichment value.
`meanpp`	The mean posterior probability of the probes in the enriched region.
`nprobe`	The number of probes in the enriched regions. nprobe = rend - rstart + 1.

Each enriched region was annotated with the gene name based on `gstart` and `gend`. These binding region of WOX5 transcription factor were overlapped with gradients lists obtained from stem cell niche cell types to analyze the involvement of the WOX5 enriched region in the increasing and decreasing gradient.

5 Results

This chapter is divided into the three subsections. The first section shows the results of differential expression analysis and In the second section, we show the results of ChIP-chip data analysis for the WOX5 transcription factor. The third section shows the results of biclustering techniques and Go analysis of the clusters.

5.1 Differential Expression Analysis

5.1.1 differentially expressed genes (DEGs)

We obtained the cell specific genes of QC, CSC and CRC by differential expression analysis. The number of up-regulated/down-regulated cell specific genes of stem cell niche cell types were obtained as follows:

- QC=10661
- CRC=9602
- CSC=4403

Each genes list were validated by Dr. Aichinger marker genes list [43]. The number of CRC genes and CSC genes are 90% and 41.3% of QC genes respectively.

The cell-specific genes were also ranked by log fold change because greater log fold change means greater gene expression in one cell type compared to another cell type which shows the enrichment level of that gene. Other reason for choosing the log fold as ranking criteria is that more reproducible results could be obtained by log fold change and it is directly measured by microarray. On other hand, p-values are more incorporate the signal-to-noise ratio [60]. The top 20 cell-specific DEGs of stem cell niche cell types are shown in Table 5, Table 6 and Table 7.

Begin of Table			
Systematic Name	Adjusted P value	Log fold change	Description
AT3G46040.1	2.84E-16	14.98	RPS15AD, regulated by TCP20.
AT5G59690.1	5.80E-14	12.12	Histone H4
AT2G25450.1	9.72E-20	11.23	Putative 2-oxoacid dependent dioxygenase, glucosinolate biosynthesis
AT5G45500.1	3.77E-16	11.132	RNI-like superfamily protein
AT5G11280.1	5.59E-20	11.07	Tail fiber;
AT5G48020.1	1.83E-19	10.52	2-oxoglutarate (2OG) and Fe(II)-dependent oxygenase superfamily protein
AT5G28920.1	4.74E-30	10.20	Hypothetical protein
AT2G40010.1	2.63E-32	10.19	Ribosomal protein L10 family protein
AT4G09310.1	6.28E-25	10.10	SPla/RYanodine receptor (SPRY) domain-containing protein; putative scaffolding and microtubule binding
AT1G35612.1	2.31E-24	10.09	pseudogene of Ulp1 protease family protein
AT3G59890.1	1.67e-22	10.07	Dihydrodipicolinate reductase; plastid lysine biosynthesis.
AT5G52420.1	1.52e-27	10.05	Unknown; putative transmembrane domain.
AT4G09200.1	1.96e-22	10.03	SPla/RYanodine receptor (SPRY) domain-containing protein.
AT2G05830.1	1.43e-26	9.95	Methylthioribose-1-phosphate isomerase; methionine salvage.
AT5G23830.1	2.731e-17	9.952	MD-2-related lipid recognition domain-containing protein; sterol transport.
AT2G18193.1	6.06e-16	9.93	AAA-ATPase; P-loop containing nucleoside triphosphate hydrolases superfamily protein.
AT5G50565.1	1.43e-26	9.967	unknown protein
AT4G20480.1	1.67e-24	9.8778	Putative endonuclease or glycosyl hydrolase.

Continuation of Table 5			
Systematic Name	Adjusted P value	Log fold change	Description
AT3G62460.1	1.11e-28	9.749	Putative endonuclease or glycosyl hydrolase.
AT5G45430.1	8.88e-20	9.71	Protein kinase superfamily protein; MAPK-like

Table 5: Top 20 significant differentially expressed genes (DEGs) of the quiescent center (QC). All genes are significant at the 0.05 and ranked by log fold change.

For the top QC genes, no particular gene classes stand out. None are previously known to be QC-enriched, except for AT3G59890 (Dihydrodipicolinate reductase), which was identified by Aichinger [43], but not confirmed. There are two oxygenases, two SPRY domain-containing proteins, two putative endonuclease or glycosyl hydrolases and a MAPK-like protein. None of these or other proteins in the list seem to have a biological pathway in common.

Begin of Table			
Systematic Name	Adjusted P value	Log fold change	Description
AT4G29200.1	6.60E-18	8.57	Beta-galactosidase related protein; root enriched
AT3G10950.1	4.66E-18	8.34	Putative RPL37AB; zinc-binding ribosomal protein family protein.
AT3G14250.1	6.81E-13	6.79	RING/U-box superfamily protein; protein ubiquitination
AT2G42840.1	7.19E-15	6.54	PDF1 (Protodermal factor 1); extracellular
AT1G70895.1	2.30E-09	6.4	CLE17 (CLAVATA3/ESR-related protein 17)

Continuation of Table 6			
Systematic Name	Adjusted P value	Log fold change	Description
AT1G63650.1	2.30E-10	6.44	EGL3, a bHLH transcription factor. Functionally redundant with GL3 and TT8 and interacts with GL3, TTG1, GL1, PAP1 and 2, CPC and TRY. Expression repressed by WER and activated by CPC/TRY.
AT5G17700.1	6.27E-08	5.75	MATE efflux family protein;
AT4G00480.1	4.10E-09	5.62	ATMYC1, a bHLH transcription factor repressing the activity of TRY and CPC through nuclear export. Regulated by GL3.
AT5G15360.1	7.35E-18	5.42	Putative transmembrane protein.
AT1G55200.1	1.16e-11	5.29	Plasma membrane kinase with adenine nucleotide alpha hydrolases-like domain-containing protein.
AT2G15790.1	3.70e-14	5.29	SQN or CYP40 (cyclophilin 40). It is specifically required for the vegetative and genetically interacts with floral meristem determinacy. Belongs to carboxylate clamp (CC)-tetratricopeptide repeat (TPR) family. Interacts with HSP90 as co-chaperone.
AT3G20880.1	3.05e-10	5.11	WIP4 zinc finger transcription factor is a paralog of NTT and along with WIP5, acts redundantly in cell fate determination during primary root development. Activated by PAN and regulated by MP.
AT1G70080.1	1.87e-14	5.01	Terpenoid synthase 6. Expressed in roots and products include dolabellane type diterpenes.
AT4G38410.1	8.24	4.98	Dehydrin family protein.
AT3G48770.1	2.03	4.88	Putative ATP/DNA binding protein.

Continuation of Table 6			
Systematic Name	Adjusted P value	Log fold change	Description
AT5G41315.1	1.87	4.81	GL3 (GLABRA 3), a bHLH transcription factor. Bound by WER to activate GL2 and CPC, which diffuses.
AT5G39220.1	6.99	4.78	Putative alpha/beta-Hydrolases superfamily protein.
AT1G11510.1	3.42	4.77	Putative DNA-binding storekeeper protein-related transcriptional regulator.
AT3G19500.1	3.48	4.74	BHLH113, a bHLH transcription factor.
AT3G53040.1	6.37e-11	4.72	late embryogenesis abundant protein, putative / LEA protein.

Table 6: Top 20 significant differentially expressed genes (DEGs) of columella stem cells (CSC). All genes are significant at the 0.05 and ranked by log fold change.

Only one gene, AT3G19500 (BHLH113, a bHLH transcription factor), in the top CSC list had been found before. Dr. Aichinger [43] had identified is as CSC-specific and confirmed it by expression analysis. In spite of that, this list is interesting because four bHLH transcription factors are present; GL3, EGL3, ATMYC1 and BHLH113. In addition, one other transcription factor (WIP4) and two putative transcription factors are in the list. Signalling is represented by PDF1, CLE17 and a putative membrane kinase.

Begin of Table			
Systematic Name	Adjusted P value	Log fold change	Description
AT5G44417.1	1.03E-12	9.93	pseudogene of FAD-binding Berberine family protein.
AT5G44440.1	6.42E-11	8.95	FAD-binding Berberine family protein, putative cell wall localization.
AT4G16215.1	2.06E-15	8.93	hypothetical protein;

Continuation of Table 7			
Systematic Name	Adjusted P value	Log fold change	Description
AT4G13230.1	5.21E-16	8.32	Late embryogenesis abundant protein (LEA) family protein.
AT5G19100.1	5.35E-17	7.75	Aspartyl protease family protein.
AT5G44410.1	5.21E-08	7.45	FAD-binding Berberine family protein, localized to cell wall.
AT3G02110.1	8.25E-11	7.31	Serine carboxypeptidase-like 25.
AT1G22880.1	4.31E-14	7.22	CEL5 (cellulase 5), endoglucanase.
AT5G14740.1	5.78E-16	7.15	BCA2 (Beta carbonic anhydrase 2, chloroplastic)
AT5G38780.1	6.32e-16	7.01	Putative S-adenosyl-L-methionine-dependent methyltransferases superfamily protein.
AT4G10490.1	3.65e-15	6.90	DLO2 (DMR6-LIKE OXYGENASE 2), 2-oxoglutarate (2OG) and Fe(II)-dependent oxygenase superfamily protein.
AT1G23210.1	8.26e-13	6.88	GH9B6 (glycosyl hydrolase 9B6), putative extracellular endoglucanase.
AT5G42600.1	1.77e-10	6.81	MRN1 (marneral synthase 1), an oxidosqualene synthase. Crucial for growth and development.
AT4G10350.1	1.62e-14	6.69	ANAC070 or BRN2 (BEARSKIN2), a NAC domain, putative transcription factor. BRN1 and BRN2, control the cell wall maturation processes that are required to detach root cap layers from the root.
AT2G23630.1	2.23e-09	6.65	SKU5 similar 16, putative redox protein of cell wall.
AT3G16450.2	5.47e-12	6.65	Mannose-binding lectin superfamily protein, localized in plasmodesma.

Continuation of Table 7			
Systematic Name	Adjusted P value	Log fold change	Description
AT5G19140.1	3.49e-16	6.58	Aluminum induced protein with YGL and LRDR motifs, localized to plasma membrane.
AT2G19590.1	7.02e-16	6.53	ACO1 (ACC oxidase 1), ethylene biosynthetic process.
AT5G24070.1	5.71e-08	6.39	Putative peroxidase superfamily protein, localized to cell wall.
AT1G07160.1	2.4e-10	6.38	Putative protein phosphatase 2C 2
AT1G54010.1	4.59e-08	6.37	Putative GDSL-motif esterase/acyltransferase/lipase, localized to ER, vacuole membrane and plasmodesma.

Table 7: Top 20 significant differentially expressed genes (DEGs) of columella root cap (CRC). All genes are significant at the 0.05 and ranked by log fold change.

The top CRC-enriched list contained one known CRC-specific gene (BRN2) and two genes (AT1G22880 and AT2G19590) from Dr. Aichinger [43], both of which were confirmed by marker expression. In contrast to the CSC list, there are few transcription factors (only BRN2), but many (13 of 20) metabolic enzymes, 2 of 20 involved in signaling, 1 of 20 carbohydrate binding, and the remaining three of unknown function. Of the enzymes there are five oxidoreductases, two cellulases and two peptidases. Remarkably, 11 of 20 proteins localize to the cell periphery; e.g., ER (1 of 11), plasma membrane (1), plasmodesma (2), extracellular (2) or to the cell wall (5). One of the 20 localizes to the plastid.

5.1.2 Gradient

In Arabidopsis thaliana, The QCs are undifferentiated cells, CSCs produce cells destined to differentiate and CRCs are most differentiated cells among these three cell types. There are a number of expressed genes in these cell types which makes decreasing expression gradient from QCs to CRCs and increasing expression gradient from QCs to CRCs [1] .

We found a number of genes from the DEGs list of QC, CSC and CRC cell types which makes gradient by log fold change. We found increasing and decreasing gradient between the DEGs of QC, CSC and CRC. The increasing gradient genes are those genes which increase in expression level from QC to CSC to CRC. Conversely, the decreasing gradient genes are those genes which decrease in expression level from QC to CSC to CRC. Tables 8 and Table 9 show the increasing gradient and decreasing gradient respectively of the DEGs.

Begin of Table		
Systematic Name	Log fold change	Description
AT5G44417	-8.54	pseudogene of FAD-binding Berberine family protein
AT5G19100	-7.94	Aspartyl protease family protien
AT5G44440	-7.87	FAD-binding Berberine family protein
AT4G13230	-7.21	Late embryogenesis abundant protein (LEA) family protein
AT3G26200	-6.04	CYP71B22 (cytochrome P450, family 71, subfamily B, polypeptide 22) oxidoreductase. The mRNA is cell-to-cell mobile.
AT5G44410	-6.03	FAD-binding Berberine family protein
AT3G02110	-5.65	Other names: SCPL25, SERINE CARBOXYPEPTIDASE-LIKE 25
AT1G50720	-5.64	Stigma-specific Stig1 family protein
AT5G19140	-5.63	aluminum induced protein with YGL and LRDR motifs
AT1G06080	-5.53	ADS1 (Delta-9 acyl-lipid desaturase 1), fatty acid biosynthesis.
AT5G26290	-5.51	MATH domain containing TRAF-like protein anchored to membrane. Mutants show defects in both male and female gametophyte development. Potentially regulated by siRNA/miR777.
AT3G48460	-5.45	GDSL-motif esterase/acyltransferase/lipase, required in fatty acid metabolism.
AT4G10490	-5.27	DLO2 (DMR6-LIKE OXYGENASE 2), 2-oxoglutarate (2OG) and Fe(II)-dependent oxygenase superfamily protein

Continuation of Table 8		
Systematic Name	Log fold change	Description
AT1G65510	-5.26	Putative transmembrane protein
AT1G52060	-5.23	Mannose-binding lectin superfamily protein
AT1G23160	-5.22	Auxin-responsive GH3 family protein
AT1G07610	-5.19	MT1C (metallothionein 1C), binds to ribosome. The mRNA is cell-to-cell mobile.
AT5G14740	-5.14	BCA2 (Beta carbonic anhydrase 2, chloroplastic)
AT5G35950	-5.14	Mannose-binding lectin superfamily protein
AT4G08555	-4.99	hypothetical protein;

Table 8: Top 20 increasing gradient genes of QC, CSC and CRC ranked by log fold change

Begin of Table		
Systematic Name	Log fold change	Description
AT1G46264	6.60	Encodes SCHIZORIZA, a member of Heat Shock Transcription Factor (Hsf) family. Functions as a nuclear factor regulating asymmetry of stem cell divisions.
AT5G63660	6.51	Predicted to encode a PR (pathogenesis-related) protein.
AT3G62760	6.20	Encodes glutathione transferase belonging to the phi class of GSTs
AT1G03840	6.08	MGP is a nuclear-localized putative transcription factor with three zinc finger domains.
AT5G10170	6.03	myo-inositol-1-phosphate synthase isoform 3.Expressed in leaf, root and silique. Immunolocaliazation experiments with an antibody recognizing MIPS1, MIPS2, and MIPS3 showed endosperm localization.

Continuation of Table 9		
Systematic Name	Log fold change	Description
AT3G50870	5.89	MNP (MONOPOLE) encodes a GATA transcriptional regulator required to position the proembryo boundary in the early embryo. Regulates shoot apical meristem and flower development.
AT2G14660	5.81	thymocyte nuclear-like protein;(source:Araport11)
AT5G28640	5.68	AN3 (ANGUSTIFOLIA3) or GIF1 (GRF1-interacting factor 1), SSXT family transcriptional regulator.
AT4G28100	5.61	Plasma membrane protein, GPI-anchored
AT3G04520	5.50	threonine aldolase 2
AT1G29270	5.40	Unknown protein.
AT2G37300	5.40	Putative transmembrane protein;(source:Araport11)
AT5G22580	5.35	Stress responsive A/B Barrel Domain-containing protein
AT2G06200	5.30	GRF6 (growth-regulating factor 6), a transcriptional regulator.
AT5G25490	5.20	Ran BP2/NZF zinc finger-like superfamily protein
AT1G72210	5.17	BHLH96 (basic helix-loop-helix) DNA-binding transcriptional regulator.
AT3G50230	5.13	Leucine-rich repeat protein kinase family protein
AT4G03210	5.03	XTH9 encodes a member of xyloglucan endotransglucosylase/hydrolases (XTHs) that catalyze the cleavage and molecular grafting of xyloglucan chains function in loosening and rearrangement of the cell wall.
AT3G56220	5.02	Putative transcriptional regulator
AT5G24900	5.00	CYP714A2 (cytochrome P450, family 714, subfamily A, polypeptide 2), putative oxidoreductase.

Table 9: Top 20 decreasing gradient genes of QC, CSC and CRC ranked by log fold change

5.2 Involvement of the WOX5 Transcription Factor

The WUSCHEL RELATED HOMEOBOX 5 (WOX5) WUS family of homeodomain transcription factors which is expressed in QC and moves in CSC and CRC in Arabidopsis thaliana. On the basis of this fact that WOX5 play a central role in the maintenance of undifferentiated. We identified the binding sites of WOX5 and its direct targets in the previously found increasing gradient and decreasing gradient. These potential targets could be related to the role of WOX5 in differentiation in QC, CSC and CRC.

Begin of Table				
Systematic Name	QC vs CSC	CSC vs CRC	LogFC	Description
AT4G10490	-2.17	-3.09	-5.27	2-oxoglutarate (2OG) and Fe(II)-dependent oxygenase superfamily protein.
AT3G11180	-2.85	-1.74	-4.60	One of 4 paralogs encoding a 2-oxoglutarate/Fe(II)-dependent oxygenases that hydroxylates JA to 12-OH-JA
AT2G30230	-3.43	-1.046	-4.47	6,7-dimethyl-8-ribityllumazine synthase.
AT5G40780	-2.14	-2.23	-4.38	Lysine histidine transporter 1.
AT1G12740	-2.43	-1.87	-4.30	Cytochrome P450, family 87, subfamily A, polypeptide 2.
AT2G30550	-2.07	-2.20	-4.28	Encodes a lipase that hydrolyzes phosphatidylcholine, glycolipids as well as triacylglycerols.
AT3G42180	-1.78	-2.42	-4.20	Exostosin family protein.
AT5G02260	-1.83	-2.08	-3.91	Member of Alpha-Expansin Gene Family.
AT5G62420	-1.56	-2.32	-3.89	NAD(P)-linked oxidoreductase superfamily protein.

Continuation of Table 10				
Systematic Name	QC vs CSC	CSC vs CRC	LogFC	Description
AT1G37130	-2.31	-1.44	-3.75	Identified as a mutant resistant to chlorate. Encodes nitrate reductase structural gene. Involved in nitrate assimilation. Has nitrate reductase activity. Up-regulated by the fungus P. indica. Binds transcription factor At2g35940. The mRNA is cell-to-cell mobile.
AT1G10480	-1.35	-2.34	-3.70	Encodes a zinc finger protein containing only a single zinc finger that acts downstream of ZFP6 in regulating trichome development by integrating GA and cytokinin signaling.
AT5G01820	-1.95	-1.53	-3.49	Encodes a CBL-interacting serine/threonine protein kinase.
AT1G13750	-1.25	-2.24	-3.491	Encodes a purple acid phosphatase whose expression is responsive to both phosphate (Pi) and phosphite (Phi) in roots.
AT1G37130	-2.17	-1.27	-3.45	Identified as a mutant resistant to chlorate. Encodes nitrate reductase structural gene. Involved in nitrate assimilation. Has nitrate reductase activity. Up-regulated by the fungus P. indica. Binds transcription factor At2g35940. The mRNA is cell-to-cell mobile.
AT3G25655	-1.24	-2.09	-3.34	Similar to Inflorescence Deficient in Abscission (IDA). Involved in floral organ abscission.
AT1G30110	-2.16	-1.14	-3.30	Encodes a ppGpp pyrophosphohydrolase.

Continuation of Table 10				
Systematic Name	QC vs CSC	CSC vs CRC	LogFC	Description
AT4G30190	-1.16	-2.13	-3.29	Belongs to the P-type ATPase superfamily of cation-transporting ATPases, pumps protons out of the cell, generating a proton gradient that drives the active transport of nutrients by proton symport. has two autoinhibitory regions within the C-terminal domain. Its plasma membrane localization is light-dependent.

Table 10: Increasing gradient DEGs bind by WOX5 Transcription factor.

Begin of Table				
Systematic Name	QC vs CSC	CSC vs CRC	Log FC	Description
AT1G62350	1.234	1.44	2.67	Pentatricopeptide repeat (PPR) superfamily protein.
AT2G04480	3.0	1.80	4.32	hypothetical protein
AT2G06200	1.57	1.67	3.49	Growth regulating factor encoding transcription activator. One of the nine members of a GRF gene family, containing nuclear targeting domain. Involved in leaf development and expressed in root, shoot and flower.
AT2G45480	1.57	1.67	3.25	Growth regulating factor encoding transcription activator. One of the nine members of a GRF gene family, containing nuclear targeting domain. Involved in leaf development.

Continuation of Table 11				
Systematic Name	QC vs CSC	CSC vs CRC	Log FC	Description
AT5G13740	2.78	1.35	4.14	Encodes ZIF1 (ZINC-INDUCED FACILITATOR1), a member of the Major Facilitator Superfamily (MFS) of membrane proteins which are found in all organisms and transport a wide range of small, organic molecules. Involved in a mechanism of Zn sequestration, possibly by transport of a Zn ligand or Zn-ligand complex into vacuoles. The mRNA is cell-to-cell mobile.

Table 11: Decreasing gradient DEGs bind by WOX5 Transcription factor

5.3 Biclustering

The Biclustering search subsets of genes having similar responses in subsets of conditions. This identifies the co-regulated groups of genes which are co-responsive to a subset of conditions irrespective of how they respond to other conditions.

Our goal with biclustering was to identify genes specifically expressed in QC, CSC and CRC. We applied PLAID and FABIA on cell specific genes of stem cell niche cell types obtained from differential expression analysis. These cell specific genes were filtered by 0.05 or 0.01 adjusted p-value cutoff. There were total 16 samples where Q163, Q164, Q061, Q062, Q063, Q064 were the replicates of CRC, JS1 JS2 JS3 JS4 were replicates of CSC and wtQC3, wtQC4, QCwt5, QCwt6, QCwt7, QCwt8 were replicates of QC. The results of PLAID and FABIA are discussed in following sections.

We applied the PLAID algorithm on gene expression data to identify the clusters which may reveal the cell-specific genes. After several combinations of parameters, PLAID identified ten clusters which were validated by Dr. Aichinger marker genes[43]. The biclustering of genes that differed at the 0.01 significance level is summarized in Table 12. What is readily apparent is that none of the clusters are exclusively validated by one of the three cell types. For example, cluster1 is validated by all CRC, none of the QC, but 8 of 9 CSC Dr. Aichinger marker genes[43]. The choice made to

assign one cluster representative of each cell type was based on which cluster fully validated that cell type and had the least proportion of validation for the other two cell types. Thus, cluster1 and not cluster3 is noted as being validated for the CRC.

Clusters	Number of genes	Number of Dr. Aichinger QC maker genes in cluster	Number of Dr. Aichinger CSC maker genes[43] in cluster	Number CRC of Dr. Aichinger maker genes[43] in cluster
cluster1***	6829	0/14	8/9	16/16
cluster2*	4643	14/14	0/9	1/16
cluster3**	5931	9/14	9/9	16/16
cluster4	5847	4/14	0/9	13/16
cluster5	4861	9/14	8/9	13/16
cluster6	4227	9/14	6/9	4/16
cluster7	6536	9/14	6/9	4/16
cluster8	6782	6/14	2/9	14/16
cluster9	6017	5/14	2/9	5/16
cluster10	9083	9/14	5/9	13/16

Table 12: Summary of biclusters of genes that differed at the 0.01 significance level identified by PLAID.
* cluster validated by QC
** cluster validated by CSC
*** cluster validated by CRC

As previously mentioned, each cell type may share the gene expression patterns that could regulate more than one cell type. Figure 14 shows that QC, CSC and CRC biclusters share in common 37 regulated genes and QC and CSC have common 3147 genes, QC and CRC have 250 common genes, whereas CSC and CRC have 1106 genes. Furthermore, QC, CSC and CRC have 1148, 1467 and 4849 exclusively regulated genes, respectively. As it can be seen that PLAID identified the clusters validated by marker genes, but clusters are larger than expected, it is possible that the algorithm groups some set of genes in the validated clusters which are not cell-specific. Therefore we used another algorithm FABIA for identification of cell-specific genes.

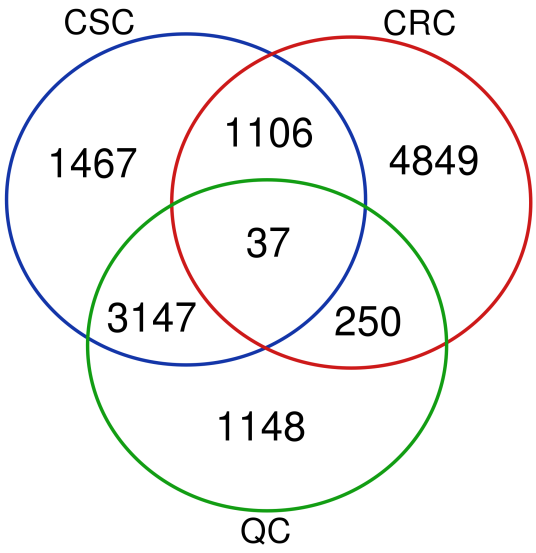


Figure 14: Venn Diagram of PLAID biclusters, validated by marker genes of QC, CSC and CRC, of genes that differed at the 0.01 significance level.

5.3.1 FABIA

We applied FABIA on our data set. The results were better in term of cluster size. The clusters validated by QC, CSC and CRC marker genes are 3155, 2811 and 2725 respectively. As Table 13 and Figure 15 shows that validated biclusters of genes that differed at the 0.01 significance level are smaller than PLAID validated biclusters. Similar to PLAID, the FABIA clusters in Table 13 are not exclusively validated by one of the cell types. The representative clusters of each of the three cell types are similarly indicated in the table. The comparison summarized in Table 14 between PLAID validated clusters and FABIA validated clusters however shows that the FABIA method gave a proportionally greater overlap between clusters.

Bicluster	Number of genes	Number of QC marker genes in cluster	Number of CSC marker genes in cluster	Number of CRC marker genes in cluster
cluster1*	3155	14/14	6/9	7/16
cluster2***	2725	7/14	0/9	16/16
cluster3	2103	9/14	0/9	15/16
cluster4**	2811	9/14	9/9	0/16
cluster5	541	2/14	6/9	0/16
cluster6	10	1/14	0/9	0/16
cluster7	116	0/14	9/9	16/16
cluster8	504	0/14	8/9	16/16
cluster9	42	4/14	0/9	4/16
cluster10	9	0/14	0/9	0/16
cluster11	7	0/14	0/9	0/16
cluster12	14	0/14	0/9	0/16
cluster13	137	1/14	0/9	0/16
cluster14	24	0/14	0/9	0/16
cluster15	126	0/14	0/9	0/16

Table 13: Summary of biclusters of genes that differed at the 0.01 significance level, identified by FABIA.
* cluster validated by QC
** cluster validated by CSC
*** cluster validated by CRC

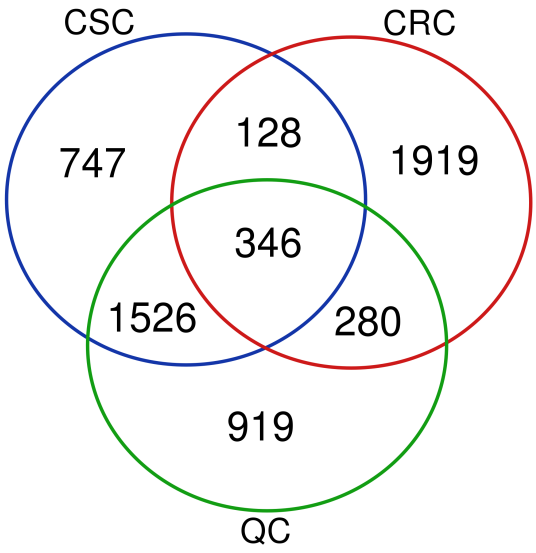


Figure 15: Venn Diagram of FABIA biclusters, validated by marker genes of QC, CSC and CRC, of genes that differed at the 0.01 significance level.

Clusters	PLAID	FABIA
Number of genes common in QC and CSC Validated clusters	3184	1872
Number of genes common in QC and CRC Validated clusters	287	626
Number of genes common in CSC and CRC Validated clusters	1143	474
Number of genes common in QC, CSC and CRC Validated	37	346
Number of genes exclusive for QC	1148	919
Number of genes exclusive for CSC	1467	1919
Number of genes exclusive for CRC	4849	747

Table 14: Comparison summary of validated clusters of PLAID and FABIA by marker genes. The genes significantly different at the 0.01 level were clustered and analyzed.

5.3.2 GO Analysis

In order to summarize the biological processes signified by these biclusters, GO analysis using the Biological Process domain was performed. The validated biclusters of genes significantly different at the 0.01 (Tables 12 and 13) and 0.05 (not shown) significance levels were subjected to GO analysis, resulting in two data sets per biclustering method. In spite of the stochastic nature of the biclustering methods used, the majority of GO terms overlapped between the 0.01 and 0.05 lists of each

method. Additionally, the 0.05 lists were not that much longer than their respective 0.01 list; therefore, the 0.05 lists of the GO Biological Process terms are presented for the PLAID (Table 15) and FABIA (Table 16) biclustering methods.

Significant PLAID GO BP Terms					
QC BP	QC Enrich	CSC BP	CSC Enrich	CRC BP	CRC Enrich
protein refolding	3.7	protein refolding	3.7	UDP-L-arabinose biosynthetic process	6.0
xylem vessel member cell differentiation	3.7	'de novo' IMP biosynthetic process	3.2	nucleotide-sugar metabolic process	5.2
arginine biosynthetic process	3.4	transcription from RNA polymerase I promoter	3.1	toxin catabolic process	4.1
'de novo' IMP biosynthetic process	3.1	mitochondrial mRNA modification	3.1	starch catabolic process	3.9
protein targeting to mitochondrion	3.1	DNA metabolic process	3.0	cell wall macromolecule catabolic process	3.7
mitochondrial mRNA modification	3.0	transcription from RNA polymerase III promoter	2.9	response to chitin	3.6
DNA metabolic process	3.0	ribosome biogenesis	2.8	jasmonic acid biosynthetic process	3.6
regulation of translational initiation	3.0	response to temperature stimulus	2.7	response to toxic substance	3.2
transcription from RNA polymerase III promoter	2.8	cytoplasmic translation	2.6	glutathione metabolic process	3.1
transcription from RNA polymerase I promoter	2.7	protein import into mitochondrial matrix	2.5	chitin catabolic process	2.9
ribosome biogenesis	2.6	mRNA transport	2.5	response to wounding	2.8
mRNA transport	2.3	RNA processing	2.4	response to karrikin	2.6

QC BP	QC Enrich	CSC BP	CSC Enrich	CRC BP	CRC Enrich
RNA secondary structure unwinding	2.1	RNA secondary structure unwinding	2.3	response to stress	2.1
chloroplast organization	2.1	chloroplast organization	2.2	ethylene-activated signaling pathway	2.0
RNA splicing	2.1	microtubule-based movement	2.1	response to oxidative stress	2.0
RNA processing	2.1	RNA splicing	2.1	protein dephosphorylation	2.0
regulation of cell cycle	2.0	DNA repair	2.0	flavonoid biosynthetic process	1.9
developmental process	1.9	covalent chromatin modification	2.0	defense response to bacterium	1.9
cell cycle	1.9	developmental process	1.9	biosynthetic process	1.8
DNA repair	1.7	regulation of cell cycle	1.9	lipid metabolic process	1.7
embryo sac development	1.7	cell cycle	1.8	defense response	1.6
cell division	1.7	embryo sac development	1.8	oxidation-reduction process	1.5
embryo development ending in seed dormancy	1.6	cellular amino acid biosynthetic process	1.8	carbohydrate metabolic process	1.4
protein folding	1.6	cell division	1.8		
response to cytokinin	1.5	embryo development ending in seed dormancy	1.7		
response to cadmium ion	1.4	protein folding	1.7		
		response to cadmium ion	1.3		

QC BP	QC Enrich	CSC BP	CSC Enrich	CRC BP	CRC Enrich

Table 15: Significant PLAID GO BP (Biological Process) terms of the columella stem cell niche. All genes significant at the 0.05 level among cell types were used for clustering.

Significant FABIA GO BP Terms					
QC BP	QC Enrich	CSC BP	CSC Enrich	CRC BP	CRC Enrich
positive regulation of secondary cell wall biogenesis	8.7	water transport	10.9	protein refolding	4.3
asymmetric cell division	5.0	regulation of root meristem growth	5.7	transcription from plastid promoter	4.3
response to toxic substance	3.0	cellular water homeostasis	4.9	tRNA methylation	4.2
response to other organism	2.8	hydrogen peroxide catabolic process	4.7	nuclear-transcribed mRNA catabolic process, exonucleolytic, 3'-5'	4.0
response to chitin	2.7	plant-type cell wall organization	4.2	amino sugar metabolic process	3.9
response to karrikin	2.6	lignin biosynthetic process	3.5	'de novo' IMP biosynthetic process	3.5
drug transmembrane transport	2.6	lipid transport	2.7	DNA metabolic process	3.3
ethylene-activated signaling pathway	2.0	transmembrane receptor protein tyrosine kinase signaling pathway	2.6	protein targeting to mitochondrion	3.2
defense response	1.6	cell differentiation	2.4	cell wall macromolecule catabolic process	3.0

QC BP	QC Enrich	CSC BP	CSC Enrich	CRC BP	CRC Enrich
transcription, DNA-templated	1.4	response to auxin	2.4	transcription from RNA polymerase I promoter	3.0
oxidation-reduction process	1.3	response to oxidative stress	2.1	ribosome biogenesis	2.7
		oxidation-reduction process	1.7	toxin catabolic process	2.5
		defense response	1.6	DNA replication	2.2
				microtubule-based movement	2.2
				chloroplast organization	2.1
				RNA secondary structure unwinding	2.1
				developmental process	2.1
				response to toxic substance	2.0
				RNA processing	2.0
				glutathione metabolic process	2.0
				embryo sac development	1.9
				cell division	1.7
				methylation	1.7
				response to salt stress	1.4

Table 16: Significant FABIA GO BP (Biological Process) terms of the columella stem cell niche. All genes significant at the 0.05 level among cell types were used for clustering.

In spite of shorter GO lists for the QC and CSC biclusters of the FABIA method, the PLAID method better matched expectations. The QC is known to have strong auxin response (DR5 marker), strong auxin synthesis (TAA1 marker) and is mitoti-

cally quiescent. The CSC also has auxin response, undergoes one mitotic event, and its daughter, the CRC, expands in length, differentiates, perhaps undergoes endoreduplication (Dolan et al. 1993), builds statoliths and thickens the cell wall. Because the CSC is mitotically active, biological processes such as DNA synthesis, transcription, ribosome assembly, translation, mitochondrial replication, plastid replication and mitosis would be expected to be enriched.

PLAID GO terms are more concordant

Inspection of the PLAID CSC results show that 21 of 27 terms are expected and 2 of 27 terms (embryo sac development and embryo development ending in seed dormancy) seem inappropriate to the CSC, but only 1 of 13 FABIA CSC terms are expected and 2 of 13 terms (lignin biosynthetic process and cell differentiation) seem inappropriate to the CSC. For the CRC, 6 of 23 PLAID and 5 of 24 FABIA terms match expectations. While no CRC PLAID terms seem inappropriate, 4 of 24 FABIA terms (DNA metabolic process, DNA replication, embryo sac development and cell division) seem inappropriate to the CRC. The QC terms fare poorly in both methods, with 1 of 26 (regulation of cell cycle) PLAID terms and 1 of 11 (asymmetric cell division) FABIA terms match expectation. Terms seemingly inappropriate to the QC; i.e., 3 of 26 (xylem vessel member cell differentiation, embryo sac development, embryo development ending in seed dormancy) from PLAID and 1 of 11 (positive regulation of secondary cell wall biogenesis) from FABIA are also present.
Because terms from the FABIA method have fewer expected and more inappropriate, it is worthwhile to describe the PLAID results in more detail. Referring to the QC column in Table 15, 19 of 26 terms are characteristic of mitotically active cells, which is unexpected because the QC is mitotically quiescent. It would be expected that QC cells are arrested in G1, and exhibit low amounts of RNA and protein synthesis. The terms (response to cytokinin and response to cadmium ion) that are neither characteristic nor inappropriate to the QC could be investigated further if necessary. Regarding the CSC, the GO terms support mitotically active cells; i.e., 2 terms DNA synthesis, 3 terms RNA processing, 2 terms mRNA, 3 terms ribosome assembly, 4 terms protein synthesis, 2 mitochondrial terms, 1 plastid term and 3 cell cycle terms. Given the large proportion of expected CSC terms, the unexpected terms (response to temperature stimulus, microtubule-based movement, response to cadmium ion and covalent chromatin modification) should be investigated in more detail. In the CRC column, 4 terms cell wall and 2 terms starch comprise the terms that would be

expected. The many unexpected terms (toxin catabolic process, response to chitin, jasmonic acid biosynthetic process, response to toxic substance, glutathione metabolic process, chitin catabolic process, response to wounding, response to karrikin, response to stress, ethylene-activated signaling pathway, response to oxidative stress, protein dephosphorylation, defense response to bacterium, lipid metabolic process, defense response and oxidation-reduction process) could be further investigated if need be.

6 Discussion

Microarray analysis can be used to obtain differentially expressed gene and gene product attributes across all condition. One of the goal of our thesis is to identify the transcritpome signature of the QC(quiescent center), CSC(Columella stem cells) and CRC (Columella cells) by using differential expressions analysis. The results of this analysis are presented in Table 5, Table 6 and Table 7. For validating these results, The results were overlapped with Dr. Aichinger maker genes [43]. These genese lists are completely validated by marker genes.

We applied another approach for validating our results. Nawy et al. [40] identified the set of 290 genes which are specifically enriched in QC and presented them by category. This set of genes was identified with all surrounding tissues and root cap. We overlapped those genes with our QC enriched genes which were obtained by comparing with CRC. We found common 241 genes out of 290. The difference of 49 genes can be explained by the difference of methods for cell types comparison. Nawy et al. [40] used linear mixed-model analysis of variance whereas our method proposed by smyth [19]. Another reason could be that they compared QC with nearly all surrounding tissues.

There are two approaches mostly used for ranking the genes either rank by adjusted p-value or fold change. The cell-specific genes which we obtained from differential expression analysis were ranked by fold change because greater log fold change means greater gene expression in one cell type compared to another cell type which shows the enrichment level of that gene. Another reason is that the more reproducible results could be obtained by fold change and it was directly measured by microarray, on other hand p-values are more incorporated in the signal-to-noise ratio [60].

When the top 20 DEGs of the QC are inspected, nothing stands out. Perhaps the reason is that these cells are mitotically quiescent, and nothing significant happens until they are reactivated. The list contains two ribosomal proteins (RPS15AD and RPL10) and Histone H4, which could indicate commitment to exiting mitotic quiescence. Perhaps the QC cells are poised to undergo mitosis if needed. Alternatively, the ribosomal proteins could be variants specific for the QC, which aid translation of

proteins required for quiescence.

In contrast, the CSC has a set of 4 functionally related bHLH transcription factors, one other transcription factor, and two putative transcription factors. One protein is related to ubiquitination, three are related to signal cascade, one is part of a chaperone complex and two could participate in synthesis and transport of a diterpene signal. Together, these indicate that gene transcription, protein regulation and signalling are required for CSC activity and production of CRC cells. The WIP4 transcription factor is involved in cell fate determination. Similarly, GL3, EGL3 and ATMYC1 together regulate cell fate through complex formation, diffusion and trancriptional regulation. Perhaps BHLH113 plays an analogous role in columella fate determination similar to GL3, EGL3, WER, TRY and ATMYC1 in trichome and root hair fate.

Rather than being a hub of transcriptional control and signalling, the top CRC DEGs are dominated by enzymes. The five oxidoreductases, two cellulases and two peptidases, plus the extracellular localization of seven of them together indicate that cell wall remodeling is important in this cell type. Indeed, the CRC has thickened cell walls compared to the CSC. Additionally, the only transcription factor in this list, BRN2, is required for CRC cell wall maturation. Four proteins (including two of the cell-wall-localized proteins): SKU5 similar 16, the lectin, the putative peroxidase and the the putative acyltransferase, localize to the plasmodesma. This localization could indicate that control of symplastic transport is important for differentiation of the CRC. Because WOX5 is known to diffuse from the QC to the CSC to prevent its differentiation, perhaps the plasmodesma between the CSC and newly formed CRC are being closed, preventing ingress of WOX5 and sealing the differentiated fate of the CRC cells.

In summary, the top 20 DEGs of the CSC and CRC are consistent with expectations, and should be investigated further by the biologist. For the QC, further refinement of its DEG list could be helpful before further interpretation. Hopefully these DEG lists provide the biologist with testable hypotheses.

In Tables 8 and 9, the DEGs are ranked by the strongest 20 increasing and decreasing gradients, respectively. The logic behind those two lists are to present genes whose products are correlated with differentiation status in the columella stem cell niche. Unsurprisingly, none of the top 20 CSC DEGs (Table 6; whose expression intensities are peaked at the CSC) are present in either of these lists. However, 9 of the top 20 CRC DEGs (Table 7) are found in the strongest 20 increasing (Table 8). In spite of this, the reverse is not true: none of the top 20 QC DEGs (Table 5) are found in the strongest 20 decreasing of Table 9. On the one hand this could mean that there

is a population of genes whose expression is tightly activated by and restricted to only the QC. On the other hand, genes that need to be upregulated for differentiation must be strongly repressed in the QC, but medium levels in the CSC do not cause differentiation until they are at highest levels in the CRC. Perhaps activation of differentiation genes is primed in the CSC so that by the next cell division, the CRC daughter promptly differentiates. The strongest 20 increasing DEGs are similar in property to the top 20 CSC DEGs: 9 of 20 are enzymes, and 10 of 20 are localized to the membrane and cell wall or extracellular region. Significantly, none of the genes are localized to the plasmodesma, which might prevent their closure between the CSC and QC, from which WOX5 diffuses from.

The strongest 20 decreasing DEGs, in contrast, have a profile similar to the top 20 CSC DEGs. There are 6 transcription regulators (SCHIZORIZA, MGP, GRF6, GIF1, GATA18 and BHLH96) and one kinase (LRR-like). The SCHIZORIZA and MGP regulate cell division resulting in different fates, which fits with the gradient along the columella stem cell niche. GRF6 and GIF1 together regulate cell proliferation and development, a property of meristematic cells. Finally, GATA18 regulates differentiation, the end products of the stem cell niche. Only 4 of 20 proteins with enzyme activity are found. Additionally, only 4 of 20 proteins are localized to the membrane and 2 of 20 to the extracellular region. Taken together, it seems that regulators of cell fate are negatively correlated with differentiation and extracellular enzymes show positive correlation. The lack of top QC DEGs in the decreasing list perhaps means that the QC cell is special among all stem and meristematic cells. Because the QC is shared among root stem cells of different lineage, one could speculate that the QC must be maintained as uncommitted to any fate until it is activated to re-enter the cell cycle to replace a stem cell. Hopefully the biologist can use this information to make experiments to understand uncommitted QC cells.

Another aim of this thesis was to find directs target of WOX5 in these three cell types. Pi et al. show that WOX5 transcription factor makes a gradient from QC to CRC and act as repressor of differentiation [1]. In light of this fact, the first two DEGs were extracted which have fold change either in increasing order (Table 8) or in decreasing order (Table 9). The WOX5 binding sites were identified by ChIP-chip analysis. The binding sites of that analysis were overlapped with DEG gradients. The 39 direct targets of WOX5 found which makes gradient in increasing order (Table 10) and 5 direct targets were found which makes gradient in decreasing order (Table 11) in these cell types. The greater number of increasing targets confirms the fact that WOX5 represses gene expression.

The increasing targets are probably repressed by WOX5, because they are lowest expressed in the QC. In the list in Table 10, 10 of the 20 are enzymes, similar to the top CRC DEGs, supporting the fact that they would be upregulated in the CRC, where no WOX5 is present. Only one transcription factor, ZFP5, is present. The ZFP5 protein is known to promote differentiation of trichomes and root hairs [61, 62], which is consistent with the need for WOX5 to repress it. Of the five decreasing targets, two are transcriptional regulators: GRF6 (also found in the decreasing DEGs Table 9), and GRF9. These genes are likely activated by WOX5, and would probably promote development of stem cells. If a biologist finds the results of cell specific genes and WOX5 direct targets interesting, these genes could be further analyzed by biological aspects.

The biclustering was used as a complementary method to identify the genes specifically expressed in QC, CSC and CRC. There were two algorithms PLAID and FABIA that were applied to the gene expression data. The gene expression data consisted of the cell DEGs below the p-value 0.01. The first algorithm applied to the gene expression data to find cell-specific genes for QC, CSC and CRC. The results of PLAID are shown in Table 12. Although results were validated by Dr. Aichinger marker genes uAichinger.However, PLAID clusters were larger than expected clusters. One of the reasons for this could be that algorithm group some set of genes in the validated clusters which do not cell specific. Therefore, FABIA were the alternative algorithm of PLAID. As Table 12 shows that FABIA gave better results in term of clusters size than PLAID. Figure 14 and Figure 15 shows that stem cell niche cell types share some genes which regulates in more than one cell type.

The GO analysis of the CSC- and CRC-validated PLAID biclusters gave the best concordance to what processes would be expected in these cell types. Even though all the Aichinger CRC genes validated in the CSC bicluster and 8 of 9 Aichinger CSC genes in the CRC bicluster, the GO terms were mostly as expected. Figure 14 shows that there is a great number (4849) of CRC genes that don't overlap with the CSC genes, explaining the agreeableness of the GO terms. However, the same figure shows a great deal of overlap (3147) genes between QC and CSC, which could explain why the GO analysis of the QC bicluster gave many terms expected for meristematic cells. To refine the GO analyses, the biologist could use only the genes exclusive for QC (1148), CSC (1467) and CRC (4849) (Table 14) as input. Not only is the enrichment in biological process domain of GO analyzed using DAVID, but other annotations were saved separately, but not further presented. These include: GO molecular function, GO cellular compartment, KEGG biological pathways, INTERPRO protein domains

and INTACT protein-protein interactions. These additional data are available upon request.

7 Conclusions

It is necessary for our understanding of stem cell niche to understand the WOX5 regulation and transcriptome signature of QC, CSC and CRC. For this purpose, we analyzed the genomic data of these cell types and offer the information about their cell specific genes and direct targets of WOX5 in the stem cell niche. Further analysis could be performed for more specific results especially in the area of biclustering. In future, it will be interesting to define more strict criteria for validation and apply other biclustering algorithms to compare their output with results of PLAID and FABIA. The DEGs and biclustering results were enigmatic for the QC, and further work would be needed to refine and interpret the gene lists. Genes contained in Tables 5 to 11 could be of great interest to biologists for further interpretation and analysis.

8 Acknowledgments

Foremost, I would like to express my sincere gratitude to my advisor Dr. Edwin Groot for the continuous support of my Master's thesis, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my thesis.

I also want to express my gratitude to the Dr. Thomas Laux and Dr. Rolf Backofen for their time, their knowledge and fruitful discussion on this work.

I will forever be thankful to my family, especially my father. He is my inspiration to do thing right in life. It would be very difficult to peruse my passion without his words of encouragement regardless of very long distance between us. I am extremely happy to be part of such a loving, understanding and supportive family.

I would like to thank Osama Ahmad and Zaid ur Rehman for proofreading and Mohsin Ali for his valuable suggestions.

Bibliography

[1] L. Pi, E. Aichinger, E. van der Graaff, C. I. Llavata-Peris, D. Weijers, L. Hennig, E. Groot, and T. Laux, "Organizer-derived wox5 signal maintains root columella stem cells through chromatin-mediated repression of cdf4 expression," *Developmental cell*, vol. 33, no. 5, pp. 576–588, 2015.

[2] E. Aichinger, N. Kornet, T. Friedrich, and T. Laux, "Plant stem cell niches," *Annual review of plant biology*, vol. 63, pp. 615–636, 2012.

[3] V. Vermeeren, S. Wenmackers, P. Wagner, and L. Michiels, "Dna sensors with diamond as a promising alternative transducer material," *Sensors*, vol. 9, no. 7, pp. 5600–5636, 2009.

[4] "Data mining and warehousing." `http://mleg.cse.sc.edu/csce822`. Accessed: 2018-05-26.

[5] S. J. Morrison, N. M. Shah, and D. J. Anderson, "Regulatory mechanisms in stem cell biology," *Cell*, vol. 88, no. 3, pp. 287–298, 1997.

[6] E. Aichinger, N. Kornet, T. Friedrich, and T. Laux, "Plant stem cell niches," *Annual review of plant biology*, vol. 63, pp. 615–636, 2012.

[7] A. Spradling, D. Drummond-Barbosa, and T. Kai, "Stem cells find their niche," *Nature*, vol. 414, no. 6859, p. 98, 2001.

[8] F. M. Watt and B. L. Hogan, "Out of eden: stem cells and their niches," *Science*, vol. 287, no. 5457, pp. 1427–1430, 2000.

[9] A. K. Sarkar, M. Luijten, S. Miyashima, M. Lenhard, T. Hashimoto, K. Nakajima, B. Scheres, R. Heidstra, and T. Laux, "Conserved factors regulate signalling in arabidopsis thaliana shoot and root stem cell organizers," *Nature*, vol. 446, no. 7137, p. 811, 2007.

[10] J. DeRisi, L. Penland, M. Bittner, P. Meltzer, M. Ray, Y. Chen, Y. Su, and J. Trent, "Use of a cdna microarray to analyse gene expression," *Nat. genet*, vol. 14, pp. 457–460, 1996.

[11] R. J. Lipshutz, S. P. Fodor, T. R. Gingeras, and D. J. Lockhart, "High density synthetic oligonucleotide arrays," *Nature genetics*, vol. 21, no. 1s, p. 20, 1999.

[12] A. Sánchez and M. de Villa, "A tutorial review of microarray data analysis," *Universitat de Barcelona*, 2008.

[13] M. E. Ritchie, J. Silver, A. Oshlack, M. Holmes, D. Diyagama, A. Holloway, and G. K. Smyth, "A comparison of background correction methods for two-colour microarrays," *Bioinformatics*, vol. 23, no. 20, pp. 2700–2707, 2007.

[14] D. Edwards, "Non-linear normalization and background correction in one-channel cdna microarray studies," *Bioinformatics*, vol. 19, no. 7, pp. 825–833, 2003.

[15] J. Quackenbush, "Microarray data normalization and transformation," *Nature genetics*, vol. 32, p. 496, 2002.

[16] Y. H. Yang, M. J. Buckley, S. Dudoit, and T. P. Speed, "Comparison of methods for image analysis on cdna microarray data," *Journal of computational and graphical statistics*, vol. 11, no. 1, pp. 108–136, 2002.

[17] T. S. Gordon Keith Smyth, "Normalization of cdna microarray data," *Methods 31, 265–273*.

[18] B. M. Bolstad, R. A. Irizarry, M. Åstrand, and T. P. Speed, "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias," *Bioinformatics*, vol. 19, no. 2, pp. 185–193, 2003.

[19] G. K. Smyth, "Linear models and empirical bayes methods for assessing differential expression in microarray experiments," *Statistical applications in genetics and molecular biology*, vol. 3, no. 1, pp. 1–25, 2004.

[20] G. K. Smyth, M. Ritchie, N. Thorne, and J. Wettenhall, "Limma: linear models for microarray data. in bioinformatics and computational biology solutions using r and bioconductor. statistics for biology and health," 2005.

[21] I. Lönnstedt and T. Speed, "Replicated microarray data," *Statistica sinica*, pp. 31–46, 2002.

[22] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proceedings of the National Academy of Sciences*, vol. 95, no. 25, pp. 14863–14868, 1998.

[23] R. R. Sokal, "A statistical method for evaluating systematic relationship," *University of Kansas science bulletin*, vol. 28, pp. 1409–1438, 1958.

[24] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[25] L. Lazzeroni and A. Owen, "Plaid models for gene expression data," *Statistica sinica*, pp. 61–86, 2002.

[26] T. Murali and S. Kasif, "Extracting conserved gene expression motifs from gene expression data," in *Biocomputing 2003*, pp. 77–88, World Scientific, 2002.

[27] A. Prelić, S. Bleuler, P. Zimmermann, A. Wille, P. Bühlmann, W. Gruissem, L. Hennig, L. Thiele, and E. Zitzler, "A systematic comparison and evaluation of biclustering methods for gene expression data," *Bioinformatics*, vol. 22, no. 9, pp. 1122–1129, 2006.

[28] Y. Kluger, H. Yu, J. Qian, and M. Gerstein, "Relationship between gene co-expression and probe localization on microarray slides," *Bmc Genomics*, vol. 4, no. 1, p. 49, 2003.

[29] Y. Cheng and G. M. Church, "Biclustering of expression data.," in *Ismb*, vol. 8, pp. 93–103, 2000.

[30] S. Hochreiter, U. Bodenhofer, M. Heusel, A. Mayr, A. Mitterecker, A. Kasim, T. Khamiakova, S. Van Sanden, D. Lin, W. Talloen, *et al.*, "Fabia: factor analysis for bicluster acquisition," *Bioinformatics*, vol. 26, no. 12, pp. 1520–1527, 2010.

[31] A. Tanay, R. Sharan, and R. Shamir, "Discovering statistically significant biclusters in gene expression data," *Bioinformatics*, vol. 18, no. suppl_1, pp. S136–S144, 2002.

[32] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *Journal of machine learning research*, vol. 5, no. Nov, pp. 1457–1469, 2004.

[33] M. Girolami, "A variational method for learning sparse and overcomplete representations," *Neural computation*, vol. 13, no. 11, pp. 2517–2532, 2001.

[34] J. Palmer, K. Kreutz-Delgado, B. D. Rao, and D. P. Wipf, "Variational em algorithms for non-gaussian latent variable models," in *Advances in neural information processing systems*, pp. 1059–1066, 2006.

[35] G. Pfundstein, *Ensemble methods for plaid bicluster algorithm.* PhD thesis, Institut für Statistik, 2010.

[36] B. Ren, F. Robert, J. J. Wyrick, O. Aparicio, E. G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, *et al.*, "Genome-wide location and function of dna binding proteins," *Science*, vol. 290, no. 5500, pp. 2306–2309, 2000.

[37] T. Sandmann, J. S. Jakobsen, and E. E. Furlong, "Chip-on-chip protocol for genome-wide analysis of transcription factor binding in drosophila melanogaster embryos," *Nature protocols*, vol. 1, no. 6, p. 2839, 2006.

[38] Q. Mo and F. Liang, "Bayesian modeling of chip-chip data through a high-order ising model," *Biometrics*, vol. 66, no. 4, pp. 1284–1294, 2010.

[39] S. M. Brady, D. A. Orlando, J.-Y. Lee, J. Y. Wang, J. Koch, J. R. Dinneny, D. Mace, U. Ohler, and P. N. Benfey, "A high-resolution root spatiotemporal map reveals dominant expression patterns," *Science*, vol. 318, no. 5851, pp. 801–806, 2007.

[40] T. Nawy, J.-Y. Lee, J. Colinas, J. Y. Wang, S. C. Thongrod, J. E. Malamy, K. Birnbaum, and P. N. Benfey, "Transcriptional profile of the arabidopsis root quiescent center," *The Plant Cell*, vol. 17, no. 7, pp. 1908–1925, 2005.

[41] M. H. Asyali, D. Colak, O. Demirkaya, and M. S. Inan, "Gene expression profile classification: a review," *Current Bioinformatics*, vol. 1, no. 1, pp. 55–73, 2006.

[42] Y. Kluger, R. Basri, J. T. Chang, and M. Gerstein, "Spectral biclustering of microarray data: coclustering genes and conditions," *Genome research*, vol. 13, no. 4, pp. 703–716, 2003.

[43] E. Aichinger, "Wox5 function in the columella stem cell niche." unpublished results, 2018.

[44] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth, "limma powers differential expression analyses for rna-sequencing and microarray studies," *Nucleic acids research*, vol. 43, no. 7, pp. e47–e47, 2015.

[45] Y. H. Yang and N. P. Thorne, "Normalization for two-color cdna microarray data," *Lecture Notes-Monograph Series*, pp. 403–418, 2003.

[46] S. Dutoit, Y. Yang, M. Callow, and T. Speed, "Statistical methods for identifying genes with differential expression in replicated cdna microarray experiments," *Stat. Sin.*

[47] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth, "limma powers differential expression analyses for RNA-sequencing and microarray studies," *Nucleic Acids Research*, vol. 43, no. 7, p. e47, 2015.

[48] G. K. Smyth, "Limma: linear models for microarray data," in *Bioinformatics and computational biology solutions using R and Bioconductor*, pp. 397–420, Springer, 2005.

[49] G. Li, Q. Ma, H. Tang, A. H. Paterson, and Y. Xu, "Qubic: a qualitative biclustering algorithm for analyses of gene expression data," *Nucleic acids research*, vol. 37, no. 15, pp. e101–e101, 2009.

[50] S. Kaiser and F. Leisch, "A toolbox for bicluster analysis in r," 2008.

[51] H. Turner, T. Bailey, and W. Krzanowski, "Improved biclustering of microarray data demonstrated through systematic performance tests," *Computational statistics & data analysis*, vol. 48, no. 2, pp. 235–254, 2005.

[52] "The database for annotation, visualization and integrated discovery (david) v6.8." `https://david.ncifcrf.gov/home.jsp`. Accessed: 2018-05-25.

[53] L. R. Huang DW, Sherman BT, "Systematic and integrative analysis of large gene lists using david bioinformatics resources," *Nature Protocols*, vol. 4, pp. 44—-57, 2009.

[54] "Gene ontology enrichment analysis and visualization tool (gorilla)." `http://cbl-gorilla.cs.technion.ac.il/`. Accessed: 2018-05-25.

[55] "Protein analysis through evolutionary relationships (panther) classification system v13.1." `http://pantherdb.org/`. Accessed: 2018-05-25.

[56] "Reduce visualize gene ontology (revigo)." `http://revigo.irb.hr/`. Accessed: 2018-05-25.

[57] S. N. S. T. Supek F, Bošnjak M, "Revigo summarizes and visualizes long lists of gene ontology terms," *PLoS ONE*, vol. 6, no. 7, p. e21800, 2011.

[58] Q. Mo, *iChip: Bayesian Modeling of ChIP-chip Data Through Hidden Ising Models*, 2012. R package version 1.32.0.

[59] M. A. Newton, A. Noueiry, D. Sarkar, and P. Ahlquist, "Detecting differential gene expression with a semiparametric hierarchical mixture method," *Biostatistics*, vol. 5, no. 2, pp. 155–176, 2004.

[60] L. Guo, E. K. Lobenhofer, C. Wang, R. Shippy, S. C. Harris, L. Zhang, N. Mei, T. Chen, D. Herman, F. M. Goodsaid, *et al.*, "Rat toxicogenomic study reveals analytical consistency across microarray platforms," *Nature biotechnology*, vol. 24, no. 9, p. 1162, 2006.

[61] Z. Zhou, L. An, L. Sun, S. Zhu, W. Xi, P. Broun, H. Yu, and Y. Gan, "Zinc finger protein5 is required for the control of trichome initiation by acting upstream of zinc finger protein8 in arabidopsis," *Plant physiology*, vol. 157, no. 2, pp. 673–682, 2011.

[62] L. An, Z. Zhou, L. Sun, A. Yan, W. Xi, N. Yu, W. Cai, X. Chen, H. Yu, J. Schiefelbein, *et al.*, "A zinc finger protein gene zfp5 integrates phytohormone signaling to control root hair development in arabidopsis," *The Plant Journal*, vol. 72, no. 3, pp. 474–490, 2012.

[63] Y. Helariutta, H. Fukaki, J. Wysocka-Diller, K. Nakajima, J. Jung, G. Sena, M.-T. Hauser, and P. N. Benfey, "The short-root gene controls radial patterning of the arabidopsis root through radial signaling," *Cell*, vol. 101, no. 5, pp. 555–567, 2000.

[64] C. A. ten Hove and R. Heidstra, "Who begets whom? plant cell fate determination by asymmetric cell division," *Current opinion in plant biology*, vol. 11, no. 1, pp. 34–41, 2008.

[65] J. Lichtenberg, A. Yilmaz, J. D. Welch, K. Kurz, X. Liang, F. Drews, K. Ecker, S. S. Lee, M. Geisler, E. Grotewold, *et al.*, "The word landscape of the non-coding segments of the arabidopsis thaliana genome," *BMC genomics*, vol. 10, no. 1, p. 463, 2009.

[66] J. Vilarrasa-Blasi, M.-P. González-García, D. Frigola, N. Fàbregas, K. G. Alexiou, N. López-Bigas, S. Rivas, A. Jauneau, J. U. Lohmann, P. N. Benfey, *et al.*, "Regulation of plant stem cell quiescence by a brassinosteroid signaling module," *Developmental cell*, vol. 30, no. 1, pp. 36–47, 2014.

[67] I. De Smet, S. Lau, U. Voß, S. Vanneste, R. Benjamins, E. H. Rademacher, A. Schlereth, B. De Rybel, V. Vassileva, W. Grunewald, *et al.*, "Bimodular auxin response controls organogenesis in arabidopsis," *Proceedings of the National Academy of Sciences*, vol. 107, no. 6, pp. 2705–2710, 2010.

[68] V. Willemsen, M. Bauch, T. Bennett, A. Campilho, H. Wolkenfelt, J. Xu, J. Haseloff, and B. Scheres, "The nac domain transcription factors fez and sombrero control the orientation of cell division plane in arabidopsis root stem cells," *Developmental cell*, vol. 15, no. 6, pp. 913–922, 2008.

9 SUPPLEMENT

AGI	resistant seedling analyzed	expression in QC/CSC/CC	Confirming, Transcriptome Data	
At5g28640	9	9	9	confirmed
At3g59890	10	0	0	not expressed
At5g02460	16	10	10	confirmed
At4g16160	12	0	0	not expressed
At3g16340	12	0	0	not expressed
At1g29790	11	10	0	not confirmed
At5g10980	12	8	8	confirmed
At4g12390	12	9	0	not confirmed
At4g16160	8	0	0	not expressed

Table 17: Confirmed genes of QC

AGI	resistant seedling analyzed	expression in QC/CSC/CC	Confirming, Transcriptome Data	
At1g05370	17	13	13	confirmed
At3g19500	18	10	10	confirmed
At3g22760	12	11	11	confirmed
At3g45210	16	13	1	not confirmed
At5g43175	17	17	17	confirmed
At5g52870	11	0	0	not expressed
At5g46310	17	11	11	confirmed
At1g16510	8	7	0	not confirmed
At1g65920	16	14	14	confirmed

Table 18: Confirmed genes of CSC

AGI	resistant seedling analyzed	expression in QC/CSC/CC	Confirming, Transcriptome Data	
At2g19590	18	11	11	confirmed
At3g10320	12	9	9	confirmed
At1g62300	12	9	9	confirmed
At1g04250	12	11	11	confirmed
At1g22880	11	10	10	confirmed
At5g66870	12	10	10	confirmed
At2g31570	13	9	9	confirmed
At5g58320	12	11	11	confirmed
At1g59870	12	9	9	confirmed
At2g47800	16	10	10	confirmed
At3g16720	15	11	11	confirmed
At1g78120	15	14	14	confirmed

Table 19: Confirmed genes of CRC

AGI	Gene	FC	adj.p-val	
AT3G54220	SCR	0,031715	2,67E-12	Wysocha-Diller et al., 2000 [63]
AT1G46264	SCZ	0,00683	9,03E-13	ten Hove et al., 2008 [64]
AT5G03150	JKD	0,074821	2,24E-13	Welch et al., 2009 [65]
AT5G17800	BRAVO	0,024593	4,57E-13	Vilarrarsa Blasi et al., 2014 [66]
AT1G33280	BRN1	15,693941	1,32E-11	Bennet et al., 2010 [67]
AT5G62165	AGL42	0,02123	1,41E-12	Nawy et al., 2005 [67]
AT1G79580	SMB	13,93322	2,69E-11	Willemsen et al., 2008 [68]
AT4G10350	BRN2	22,882477	1,83E-12	Bennet et al., 2010 [67]
AT4G30080	ARF16	13,547534	3,34E-13	Wang et al., 2005 [?]

Table 20: Confirmed genes of CRC / QC