UPPSALA
UNIVERSITET

Regional variation in gene expression in *Capsella bursa-pastoris*

Sara Kurland

Degree project in biology, Bachelor of science, 2013
Examensarbete i biologi 15 hp till kandidatexamen, 2013
Biology Education Centre
Supervisor: Karl Holm

UPPSALA
UNIVERSITET

Abstract

Historical instances of local adaption and demography may affect variations in patterns of gene expression. Studying these patterns may shed light on a species' or populations' evolutionary history. This study focuses on a world-wide sample of the wide-spread weed *Capsella bursa-pastoris* in order to analyze variations in patterns of gene expression, which may reflect a region-wise division of the data set, made according to previous studies observing differences between the two regions putatively caused by local adaptations or demographic history. This was done by examining differentially expressed genes from 24 separate individuals of *C. bursa-pastoris* collected from North America, Western Eurasia and Asia. Differentially expressed genes were found in contrasts between the two regions in binomial tests, as well as in contrasts along a latitudinal and longitudinal cline within each region. The region-wise contrast resulted in 2,328 differentially expressed genes. Latitudinal contrasts resulted in 620 differentially expressed genes in the first region and 3 genes in the second. The results from the longitudinal contrast yielded 4 and 94 genes for each region, respectively. This study shows that the difference in gene expression between regions is greater than the variation within each region and that this division is reflected in the amount of differentially expressed genes.

Contents

Introduction

Natural selection cannot occur without genetic variation. Both demographic history and selection create genetic variation and may contribute to the formation of populations and to instances of speciation. Patterns of variation in gene expression may therefore be described as an intermediate phenotype between these different levels of variation. Studying patterns of variation in gene expression in a species or population may thereby aid in revealing its' evolutionary history.

Events that may alter a species' or populations' evolution are demographic history and local adaptation. The latter is essential to all organisms as they are governed by external factors, *e.g.* the climate they live in, and must acclimatize to their surroundings in order to survive. This is of particular importance for immobile creatures such as plants. An external factor that is essential to a plants' life is light. It is not only crucial in daily processes related to plant growth and survival, but also functions as a way to measure seasonal changes. This is of importance since the quality of light, as well as the relative lengths of dark and light periods, varies as seasons change. Light also varies along a latitudinal cline. Local adaptions in widespread plant species may therefore be expected to vary with respect to latitude, and so also patterns of genetic expression in genes regulating photoperiodic responses. Hence, the presence of latitudinal clines in gene expression is often treated as signs of local adaptions (Mitchell-Olds and Schmitt 2006). The presence of longitudinal clines on the other hand is sometimes regarded as results of random changes caused by population structure (Mitchell-Olds and Schmitt 2006).

Capsella bursa-pastoris

This study focuses on the herbaceous weed *Capsella bursa-pastoris*. The small genus *Capsella* consists of four species; *C. rubella, C. grandiflora, C. bursa-pastoris* and the newly found *C.orientalis* (Hurka *et al.* 2011). They are all diploid with 2n = 2x = 16 chromosomes, with the exception of *C. bursa-pastoris,* a tetraploid with 2n = 4x = 32 chromosomes (Hurka *et al.* 2011). The latter is an annual, selfing, long-day plant. It is the fifth most common flowering plant in the world and can be found on virtually all continents, except for the Antarctic. Although its origin is yet unknown it has been estimated to an area around the eastern Mediterranean and Middle East (Ceptilis and Lascoux, 2005). It has previously been shown to have a small effective population size, indicating rapid expansion (Ceptilis and Lascoux, 2005). Little is known of its' demographic history, although certain studies suggest it has spread in multiple events, most of which coincide with human patterns of dispersal (Ceptilis and Lascoux, 2005). Today, *C. bursa-pastoris* is spread by humans traveling across the globe, which not only introduces the species to a wide variety of climates, but also contributes to the way the species is spread and thereby alters the population structure. The species' high level of phenotypic variability may therefore not only be explained by considering it being subject to a wide range of climates, as discussed above, but also by its patterns of dispersal (Slotte *et al.* 2008). Consequently, *C. bursa-pastoris* constitutes an interesting organism to study with respect to patterns of variation in gene expression due to local adaptations and historical demographic events. Less is known of the regional variations at the molecular level, making the species further relevant for this project.

Research aims

This study analyzes differentially expressed genes from 24 separate individuals representing an equal amount of populations, ranging from Europe to Western Eurasia, China and North America. The data set was divided into two regions, merited by previous results indicating differentiation in genetic and phenotypic variation between the two (Slotte 2009, Holm 2010). Due to the putative dissimilarities in demographic and evolutionary history, which each region represents, differences in patterns of gene expression between the two regions are to be expected. This study examines whether the regional division is reflected in differences in patterns of gene expression between the two regions. Variations in patterns of gene expression are also tested within each region. The samples used were previously collected at various locations. Differences with respect to a latitudinal cline can be expected due to local adaptations, while differences concerning a longitudinal cline are perhaps better explained by demographic history (Mitchell-Olds and Schmitt 2006). Thus, differences can be expected between, as well as within the two regions. In summary, this study aims to analyze variations in patterns of gene expression in *C. bursa-pastoris* which may reflect the division of the data set into two regions, and that may be a consequence of local adaptations or demographic history.

Methods and materials

Plant material

Seed samples were previously collected from 24 individuals of *C. bursa-pastoris* (Brassicaceae) representing the same number of populations from Europe, North America, North Africa, the Middle East, Russia, China and Taiwan. The samples were subdivided into two regions, each compromising 12 populations (Fig. 1) (Holm 2010). The first region, region I, comprised the samples from Western Eurasia and North America. The latitudinal cline herein stretched from 63° to 32°, a difference of 31°. The longitudinal cline ranged from 131° to 97°, covering 34°. The second region, region II, included the samples from Eastern Asia. The latitudinal cline included here encompassed 21° (from 45° to 24°) and the longitudinal cline 24° (from 126° to 101°). Besides containing narrower clines than those included in region I, region II exhibits a larger degree of variability in local climatic conditions. For example, most sample localities in region I were at sea level, while samples in region II were collected from sea level up to altitudes of 1500 m (Holm 2010). This variation, as well as the comparatively narrow latitudinal range covered in region II, may decrease the possibility of detecting significant clinal variation in the region. Region II may therefore be expected to exhibit a slighter variation in gene expression compared to region I.

In order to facilitate germination, the seeds were stratified for four days on moist filter paper at 4°C. The seedlings were transferred to soil pots at randomized positions in a growth chamber with a long day photoperiod (16h light and 8h dark) in 20 °C. Plant tissues were collected from whole seedlings after approximately two weeks, at ZT 8 (8 hours into the light-period at mid-day). Ploidy levels were determined by flow cell cytometry on leaf tissue at Plant Cytometry Services (Schijndel, the Netherlands) in order to ensure that the collected samples comprised *C. bursa pastoris* and not the morphologically similar *C. rubella*.

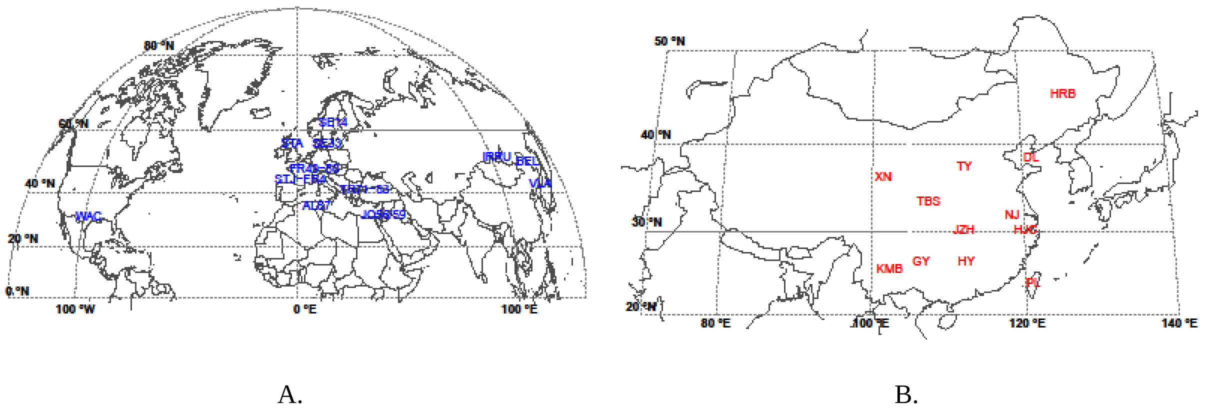


Figure 1. Geographical distribution of *C. bursa-pastoris* populations from which the transcriptome was

sequenced in region I (A). Geographical distribution of *C. bursa-pastoris* populations from which the transcriptome was sequenced in region II (B).

Sequence data

mRNA was previously isolated from the sampled seedlings. Sequence data was generated from the transcriptome by high-throughput sequencing (Illumina RNA-Seq), a technology in which vast amounts of short reads are produced in parallel by performing paired-end sequencing in a "forward-reverse" orientation, thus producing two reads per sequence fragment (Strickler *et al*. 2012). The reads were first filtered as to only include complete reads, and then mapped to a reference genome from the closely related *C. rubella* which encompasses 26,251 loci across which sequencing generated 1,730,362,730 reads, comprising 35,000,000-100,000,000 reads per sample. Mapping of *C. bursa-pastoris* to *C. rubella* resulted in 22,300 loci across which 1,480,748,034 reads (i.e. 740,374,017 counts) were generated. Lastly, the denoted loci were annotated by comparison with homologous locus tags in *Arabidopsis thaliana* and GO annotation run. This contrast was assumed appropriate since despite being separate species, *C. rubella* (a very close relative of *C. bursa-pastoris*) and *A. thaliana* are closely related (Slotte *et al.* 2007, Hurka *et al.* 2011). In fact, the genus Capsella is one of closest wild relatives of Arabidopsis (Hurka *et al.* 2011, Huang *et al.* 2012). Furthermore, comparative mapping studies have shown that but with a few exceptions, there is near to complete conservation of gene order and content between the *Capsella* genus and *A. thaliana* (Slotte *et al.* 2007).

Differential expression

The count data was used as input files for the software program Differential Expression analyses for Sequence count data (DESeq) (Bioconductor 2004, Anders and Huber 2010), which was run in the statistic environment R 2.15.2 (R Core Team 2012), in order to study gene expression. The parametric DESeq algorithm uses negative binomial distributions to test for differential expression in gene counts by assuming a local linear relationship between over-dispersions and levels of mean expression of the data. The data quality was assessed by sample clustering and visualization in heatmaps and PCA-plots, with the argument method="blind" when estimating the sample-to-sample distances (also known as dispersions). This setting generates a parametric version of the data where zero count values as well as non-zero values are taken into consideration in a variance stabilizing transformation. This transformation method is justified because it moderates the variability, which is generated when (logarithmic) fold change estimates are performed, in particular for small counts. The dispersions around a gene have namely been found to depend on the amount of overall counts, where the standard deviation is high for low count values and causes only very high fold changes to be called on as significant, which may drown informative signals in the remaining data and result in false rejections (Anders & Huber 2010).

Beside depending on dispersion, the variance around a gene depends on the uncertainty in measuring concentrations of the numbers of gene counts (shot noise). Heatmaps and PCA-plots generate parametric estimations of the dispersions around genes in a count matrix in order to illustrate similarities and dissimilarities between samples (Anders & Huber 2010). Hence, data points (samples) that might prove to be detrimental to the study, or otherwise unexpectedly divergent, may be discovered and consequently removed. This visualization may also aid in assessing the suitability of both the tests to be performed as well as the experimental design, since the significance of shot noise depends on the level of gene expression.

Another quality control of the test was done by variance estimation, and may be seen in the supplementary data (Fig. S1). Since the dispersion is greater than the shot noise in highly expressed genes, while the reverse relationship is true in lowly expressed genes, plotting dispersions against the mean value of gene expression (sampling variance against “true” gene variance) illustrates how well the data accord to the expectations of our approach, and thus enables a quality assessment of the data. The results regarding the contrasts run in this study II show a relatively high sample-to-sample variance (Fig. S1, S2, S3).

Differentially expressed genes were found between the conditions by performing a binomial test, in which significant deviations from theoretical expectations concerning the distribution of observations from two contrasting conditions is tested. This was done with the sharing mode set to “maximum” and fit type to “local”, as instructed by the manual (Anders and Huber 2010). The former argument uses a local fit instead of a parametric fit. The latter argument adjusts for the variance created, in particular by low count values. As previously mentioned the overall amounts of dispersions are considered in the differential expression inference. More specifically, this function assigns each gene a dispersion value by first approximating the dispersions for each gene, and then fitting a normalized curve through these estimates. The variance around the regression line reflects both sampling variance, and underlying variance between genes. The former is assumed to be represented by the per-gene estimates that lie below the line. Hence, these values are shifted upwards to be replaced by the values predicted by the line. The per-gene estimates found above the line are however assumed to represent the true variance between genes and are thus not adjusted for. The option “sharing mode” adjusts this function, and of the three possible settings “maximum” was found to be the most conservative (the amount of differentially expressed genes generated when contrasting region I and II from “gene-est-only”, “fit-only” and “maximum” were 6,991, 4,580 and 2,937 respectively).

The binomial tests performed may be divided into three types of contrasts. First, differences between the two regions were compared over the complete data set. The latter two tests contrasted latitude and longitude within each region respectively. This was done by contrasting northern and southern samples in one test, and western and eastern samples in another, as defined by the latitudinal/longitudinal degree at which each sample was gathered. Four of the total twelve samples in each region were defined as latitudinal/longitudinal extremes (N/S or W/E). The remaining four samples were designated intermediate values. The regions were thus subdivided into three categories (north/intermediate/south versus west/intermediate/east), all of which contained four samples. Binomial tests were done excluding the four intermediate values so as to only contrast the extremes. In summary, binomial tests were performed within and between regions; contrasting the two regions, as well as contrasting latitudinal extremes and longitudinal extremes within each region respectively.

In order to study the biological functions of the differentially expressed genes, the obtained gene lists were compared to annotated gene lists from *A. thaliana*. This was done in the Database for Annotation, Visualization, and Integrated Discovery 6.7 (DAVID), a web-based program that analyses lists of genes and facilitates functional biological annotations (notes) by adding annotations from public genomic resources to the submitted gene list and then comparing specific functional categories in the submitted list to the complete reference genome through enrichment analysis (Glynn *et al.* 2003). Enrichment analysis is based on the assumption that similar annotations are found in similar gene members, located at similar two-dimensional placements. These genes can thus be clustered into groups of similar biological meaning and iteratively merged into clusters depending on the levels of linkage between the genes. The higher the enrichment score a cluster has the closer the linkage is between genes included and thereby the more similar biological functions they exhibit. Thus, clustering patterns are used in order to group genes into clusters of related gene members, which reflect similar biological functions.

The software Empirical Bayesian analysis of patterns of differential expression in count data (baySeq) was used to test for differentially expressed genes when contrasting region I and II in order to test the consistency of the statistical algorithms used (Hardcastle and Kelly 2010). Like, DESeq, baySeq uses a parametric approach. It however estimates posterior probabilities of previously defined models for each count, or locus, in the complete data set. This is done by considering a distribution for each count by a set of underlying parameters with a posterior distribution. The estimation of the data is subsequently compared to these distributions, enabling an assessment of the likelihood of a model. In this way, the biological patterns likely to occur in the data are studied.

Results

Differentially expressed genes between regions I and II.

Visualization of the dataset shows that the subdivision of the dataset into two regions is reasonable (Fig.2). Clustering analyses separates the complete dataset into two region wise clusters (region I and II), but for one exception, namely the Chinese individual HRB which interrupts the line of samples from region I in the heatmap (Fig. 2A) and clusters with samples from region I in the PCA-plot (Fig. 2B). This would imply that it is more similar to the samples from region I than from region II. Although the sample is geographically close to the Russian samples in region I, it does not cluster particularly close to them (Fig. 1). One possible explanation to these observations is that the sample in reality comes from region I, but has mistakenly been marked as belonging to region II. Given how the species spread, another plausible explanation could be that it belongs to region II but represents a recently introduced sample originating from region I. Due to the uncertainty of determining the sample's origin from overall gene expression patterns, HRB was omitted from all further tests. This decision was first examined by running a binomial test in which both regions were compared but with the sample HRB defined as belonging to region I instead of II, which resulted in 2,270 differentially expressed genes. 2,181 significant genes were found when performing the same test, but with HRB defined as region II. The highest number found (2,328 differentially expressed genes) occurred when HRB was omitted completely.

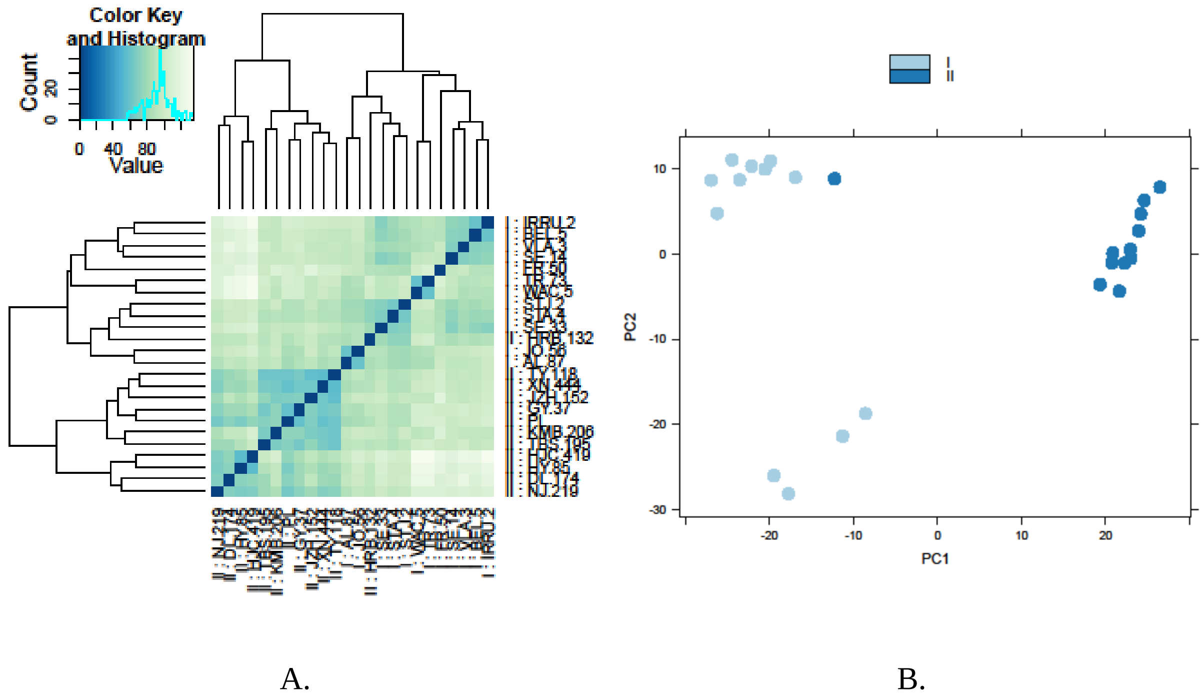


Figure 2. Heatmap over all samples with name of individual sample and region specified (A). PCA-plot over all samples from both regions contrasting region I with region II (B).

Of the differentially expressed genes found when comparing region I and II, 1,927 were successfully paired with homologous genes found in *A. thaliana* in DAVID. The five first clusters with the highest enrichment scores (3.6- 6.1) were but for one exception, mostly related to house-holding functions (Table 1). One cluster, the second contained genes related to response to radiation and light (Table 1). These genes were also found in clusters with lower enrichment scores, but which contained genes associated with photoperiodism. 198 were related to response to light as well as the distinction between red and far-red light, and distributed in 4 separate clusters. 6 genes had functions related to photoperiodism, 6 to photo reduction, and 9 to circadian rhythm (Table 2). The remaining genes (not presented) were involved in a variety of functions such as house-keeping, regulation of developmental stages, cell- and plant growth. The differences in gene expression between the two regions mainly provide genes with functions seemingly without any biological relationship, which implies that random differences have occurred, most likely due to demographic events.

Table 1. Results from DAVID representing the five clusters with the highest enrichment scores in gene lists generated from contrasting region I and II.

Annotation Cluster 1	Enrichment Score: 6,1	
	Term	**Count***

	chloroplast	354
	plastid	359
	plastid part	137
	chloroplast part	132
	plastid envelope	67
	chloroplast envelope	64
	envelope	84
	plastid stroma	54
	organelle envelope	83
	chloroplast stroma	51
Annotation Cluster 2	**Enrichment Score: 4,1**	
	Term	**Count**
	response to radiation	71
	response to light stimulus	67
	response to abiotic stimulus	140
	response to red or far red light	27
Annotation Cluster 3	**Enrichment Score: 4,061**	
	Term	**Count**
	external encapsulating structure	80
	cell wall	78
	plant-type cell wall	40
Annotation Cluster 4	**Enrichment Score: 3,7**	
	Term	**Count**
	transit peptide	107
	chloroplast	88
	plastid	85
	transit peptide:Chloroplast	80
Annotation Cluster 5	**Enrichment Score: 3,6**	
	Term	**Count**
	glycosyltransferase	40
	glucosyltransferase activity	26
	UDP-glucosyltransferase activity	23
	IPR002213:UDP-glucuronosyl/UDP-glucosyltransferase	19

*Count values presented in the table designate the number of genes included in the corresponding term. The same genes may occur in several clusters and terms and hence be counted several times.

Table 2. Homologous genes related to photoperiodic response in *A. thaliana* , differentially expressed when contrasting region I and II over both regions. Locus tags and gene names are specified, as well at their regulation in region I compared to region II.

Photoperiodism			
Enrichment score = 2.08			
Locus tag	**Gene Name**	**Log2 Fold change**	**Regulation**
AT1G25560	AP2/ERF and B3 domain-containing transcription repressor TEM1	-1.024	Down
AT2G46830	AT2G46830	-1.794	Down
AT4G35090	Catalase-2; Catalase	-1.372	Down
AT4G40060	Homeobox-leucine zipper protein ATHB-16	-0.527	Down

AT5G01040	Laccase-8	-1.991	Down
AT2G02760	Ubiquitin-conjugating enzyme E2 2	-0.896	Down
	Circadian rhythm **Enrichment score = 1.88**		
Locus tag	**Gene Name**	**Log2 Fold change**	**Regulation**
AT2G46830	AT2G46830	-1.794	Down
AT5G02840	AT5G02840	-1.331	Down
AT5G37260	AT5G37260	-2.107	Down
AT1G68830	Serine/threonine-protein kinase SNT7, chloroplastic	-0.596	Down
AT3G57040	Two-component response regulator ARR9	-0.520	Down
AT5G60100	Two-component response regulator-like APRR3	-0.959	Down
AT2G46790	Two-component response regulator-like APRR9	-0.616	Down
AT2G46830	AT2G46830	-1.794	Down
AT4G16250	Phytochrome D	-1.446	Down
	Photoreduction **Enrichment score = 0.71**		
Locus tag	**Gene Name**	**Log2 Fold change**	**Regulation**
AT2G26670	AT2G26670	-0.426	Down
AT4G16250	Phytochrome D	-1.446	Down
AT1G53090	Protein SPA1-RELATED 4	-0.609	Down
AT1G02340	Transcription factor HFR1	-0.815	Down
AT2G43010	Transcription factor PIF4	-0.380	Down
AT5G61270	Transcription factor PIF7	-0.619	Down

All of the genes related to photoperiodic response were down-regulated in region I as compared to region II (Table 2), suggesting that gene expression in these genes differ between the two regions. However, results from the binomial test as well as gene regulation and the amount of unexpressed genes indicate a significant difference in gene expression in the two regions.

Differentially expressed genes between latitudinal extremes within regions: region I

The cluster analysis shows a clear differentiation between northern and southern samples (Fig. 3). The northern samples cluster with the intermediate samples to form one cluster while the southern samples divide into yet two sub-clusters. These consist of two pairs of individuals from Algeria (AL.87), Jordan (JO.56) and Texas (WAC.5), Turkey (TR.73) respectively (Fig. 3A). The branch lengths to both sub-clusters are equal, indicating that the differences within the two are comparable; yet, the geographic distance between the latter pair is vastly greater than the former. Furthermore, the geographical distance between the three European samples is small, and the distance from these to the American relatively equal. This is not reflected in the clustering patterns, suggesting that geographical distance lacks any determining influence on clustering patterns within this region.

A similar observation may be seen when considering the northern samples. The French sample distinguishes itself from all other samples included in the north/intermediate cluster, although some samples included in the cluster exhibit greater geographic variation relative each other than relative the French sample. The two Swedish samples (SE. 33 and SE. 14) also prove interesting in that they don’t cluster together, as would be expected if geographical distance alone were influencing expression patterns (Fig. 3A).

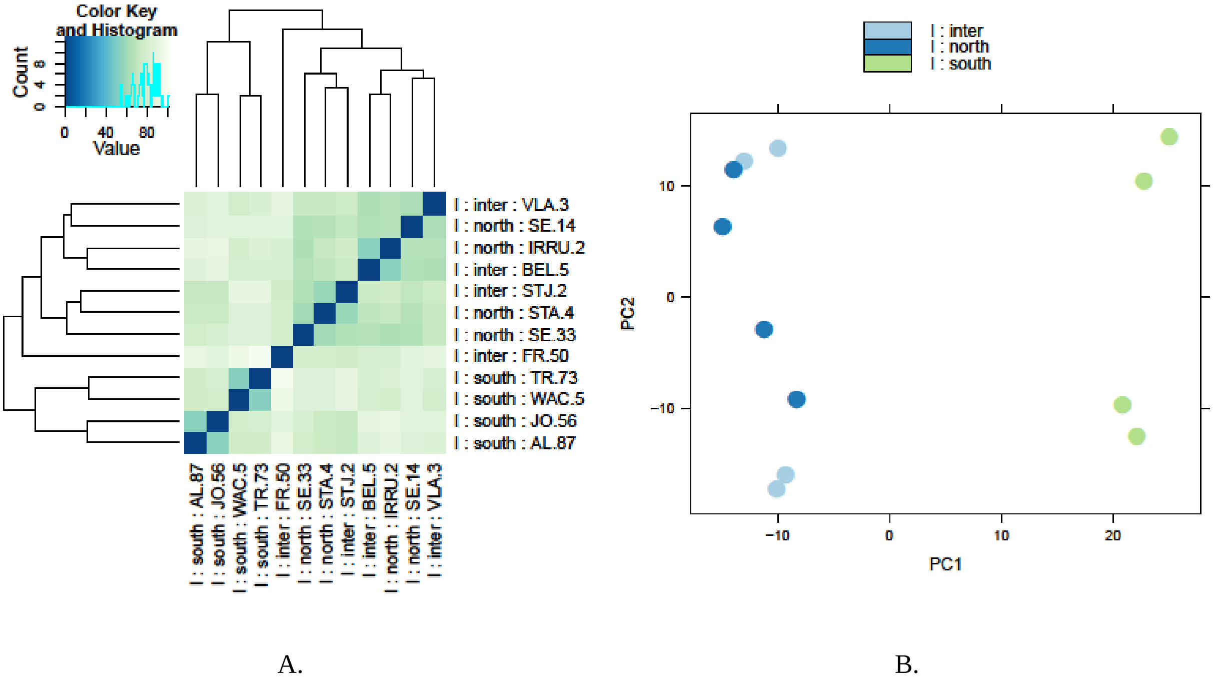


Figure 3. Heatmap over all samples in region I with name of individual sample and latitudinal definition specified (A). PCA-plot over all samples from region I with latitudinal definition specified (B).

Binomial testing, in which northern and southern extremes (intermediate values excluded) were contrasted within region I, resulted in 620 differentially expressed genes. The first five clusters with highest enrichment scores were related to various biological functions (Table 3). Their enrichment scores ranged from 2.84 to 1.83, and were thus not as high as the five clusters generated when contrasting the two regions which ranged from 6.1 to 3.6 (Table 1).

The second cluster, with enrichment score 1.970, contained genes related circadian rhythm; AT2G46830, Two-component response regulator-like APRR9, AT5G37260, AT3G46640, Adagio protein 3, Two-component response regulator-like APRR5 and Zinc finger protein

CONSTANS-LIKE 1. Three of the seven genes related to circadian rhythm in region I were up regulated and four down regulated among northern samples compared to southern, suggesting a similar pattern of expression in genes related to circadian rhythm among northern and southern samples (Table 4).

19 of the differentially expressed genes were related to light perception, and radiation, as well as the distinction between red and far-red light (Table 4). The majority of these genes, 12, were up regulated in northern samples compared to southern, while 7 were down regulated. This difference suggests a variation in patterns of expression in genes related to light perception among northern and southern samples. This may in turn indicate a correlation between variations in gene expression and the latitudinal cline, possibly caused by different local adaptations within this region.

Table 3. Results from DAVID representing the five clusters with the highest enrichment scores in gene lists generated from contrasting latitude in region I.

Term	Count*
Annotation cluster 1 Enrichment Score: 2.84	
SM00336:BBOX	7
IPR000315:Zinc finger, B-box	7
zinc finger region:B box-type 2; atypical	3
zinc finger region:B box-type 1; atypical	3
Annotation cluster 2 Enrichment Score: 1.97	
circadian rhythm	7
rhythmic process	7
Circadian rhythm	4
regulation of flower development	4
regulation of post-embryonic development	4
Annotation cluster 3 Enrichment Score: 1.91	
SM00579:FBD	11
repeat:LRR 3	15
IPR013101:Leucine-rich repeat 2	11
leucine-rich repeat	22
IPR006566:FBD-like	11
repeat:LRR 2	16
repeat:LRR 1	16
SM00256:FBOX	21
domain:F-box	22
domain:FBD	5
IPR001810:Cyclin-like F-box	21
IPR013596:FBD	6
repeat:LRR 4	8
PIRSF016997:hypothetical protein, Arabidopsis thaliana F17J16.30 type	3
Annotation cluster 4 Enrichment Score: 1.86	
glycoprotein	35
disulfide bond	28
signal	46
glycosylation site:N-linked (GlcNAc...)	34
signal peptide	46
disulfide bond	25
Secreted	27
GO:0005576~extracellular region	30

Annotation cluster 5 Enrichment Score: 1.83	
aminoglycan metabolic process	5
Glyco_18	3
chitin metabolic process	4
aminoglycan catabolic process	4
chitin catabolic process	4
GO:0005976~polysaccharide metabolic process	11
IPR001223:Glycoside hydrolase, family 18, catalytic domain	3
IPR011583:Chitinase II	3
GO:0004568~chitinase activity	4
GO:0000272~polysaccharide catabolic process	6
IPR013781:Glycoside hydrolase, subgroup, catalytic core	9
GO:0016052~carbohydrate catabolic process	8

*Count values presented in the table designate the number of genes included in the corresponding term. The same genes may occur in several clusters and terms and hence be counted several times.

Table 4. Homologous genes concerning light perception and circadian rhythm in *A. thaliana*, differentially expressed when contrasting latitudinal extremes (north and south) within region I. Locus tags, gene names and regulation specified.

Light perception, distinction between red and far-red light, radiation				**Circadian rhythm**			
Locus tag	**Gene name**	**Log2fold change**	**Regulated in north versus south**	**Locus tag**	**Gene name**	**Log2fold change**	**Regulated in north versus south**
AT2G42870	AT2G42870	1.194	Up	AT2G46830	AT2G46830	-1.961	Down
AT2G46830	AT2G46830	-1.96	Down	AT2G46790	Two-component response regulator-like APRR9	-1.185	Down
AT3G21890	AT3G21890	-2.549	Down	AT5G37260	AT5G37260	-2.282	Down
AT4G03400	AT4G03400	0.878	Up	AT3G46640	AT3G46640	1.758	Up
AT4G14690	AT4G14690	-2.068	Down	AT1G68050	Adagio protein 3	2.062	Up
AT4G15480	AT4G15480	1.639	Up	AT5G24470	Two-component response regulator-like APRR5	1.489	Up
AT5G24120	AT5G24120	-1.013	Down	AT5G15850	Zinc finger protein CONSTANS-LIKE 1	-1.507	Down
AT5G59920	AT5G59920	-4.350	Down				
AT5G63600	AT5G63600	1.282	Up				
AT1G68050	Adagio protein 3	2.062	Up				

AT3G15540	Auxin-responsive protein IAA19	1.240	Up
AT4G31500	Cytochrome P450 83B1	-3.886	Down
AT5G05690	Cytochrome P450 90A1	0.949	Up
AT1G03190	DNA repair helicase UVH6	4.717	Up
AT1G79440	Succinate-semialdehyde dehydrogenase, mitochondrial	0.813	Up
AT5G24470	Two-component response regulator-like APRR5	1.489	Up
AT2G46790	Two-component response regulator-like APRR9	-1.185	Down
AT2G06850	Xyloglucan endotransglucosylase/hydrolase protein 4	1.406	Up

Differentially expressed genes between latitudinal extremes within regions: region II

Since sample HRB was omitted, the clustering analysis only included seven samples, of which three were defined as intermediate (instead of the original four). Cluster analysis in this case shows that the Chinese sample TY 118 as a recluse (Fig 4). The remaining samples grouped together without any apparent pattern with respect to latitudinal origin.

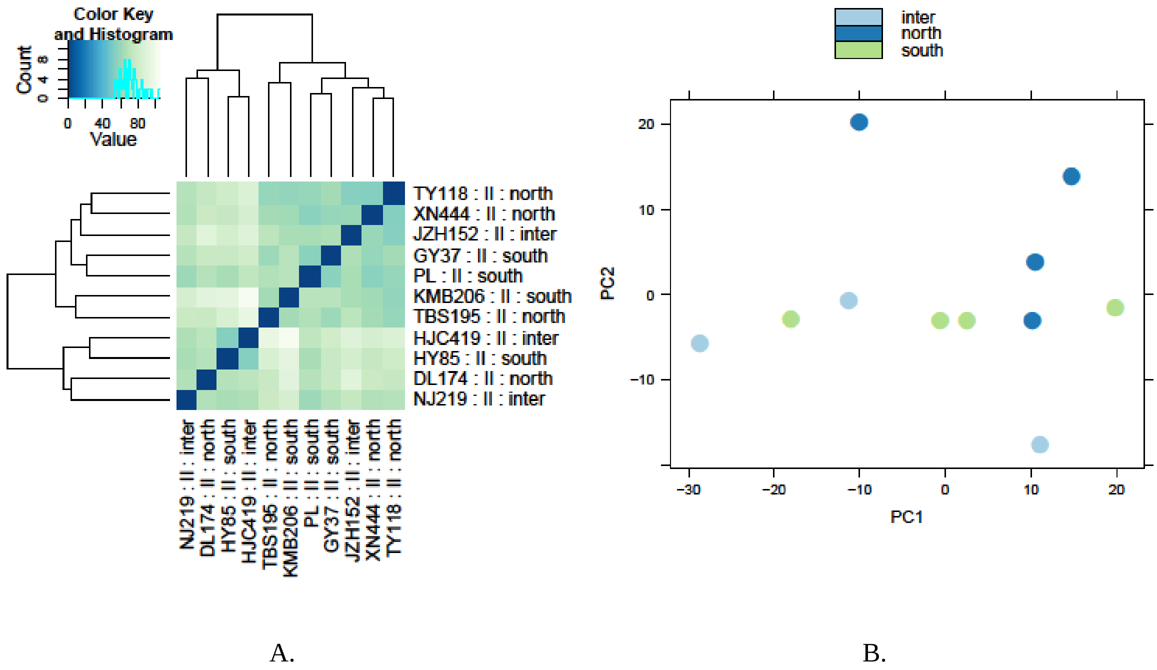
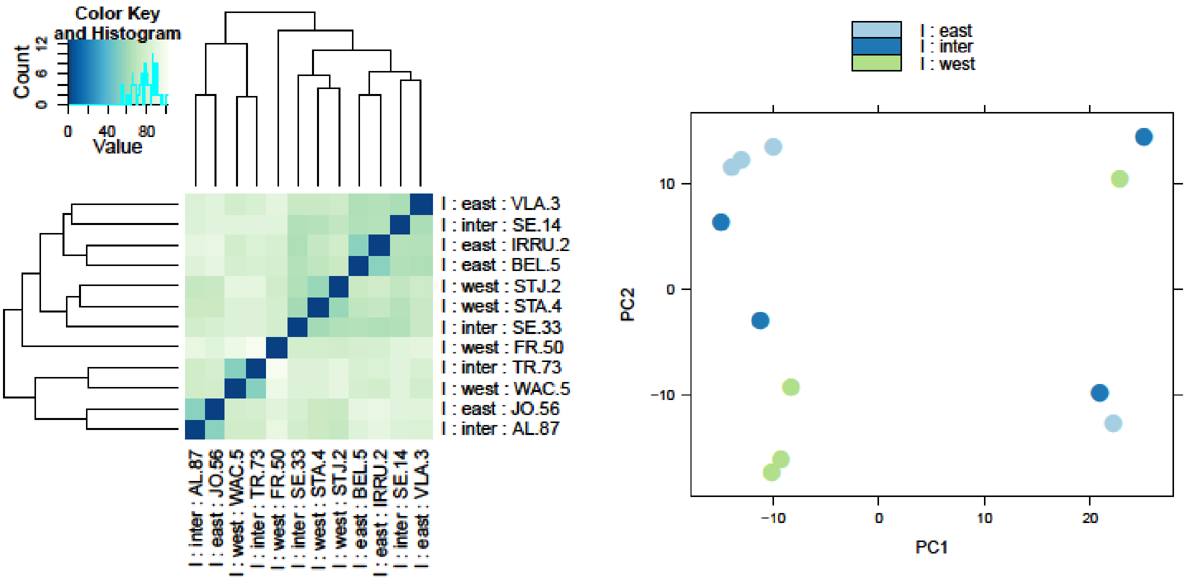


Figure 4. Heatmap over all samples in region II excluding sample HRB, with name of individual sample and latitudinal definition specified (A). PCA-plot over all samples from region II with HRB sample omitted, contrasting latitudinal definitions north versus south (B).

As might be expected from the results from clustering analysis above, binomial testing contrasting latitude within region II reported fewer genes than for the corresponding result for region I. 3 genes were reported as differentially expressed, of which two were found in *A. thaliana*, namely AT5G40595 and AT5G45220, both concerned with immune response and defense mechanisms. The former had a log2 Fold value of 6.213 and was up regulated in north compared to south. The latter was down regulated and exhibited a value of -8.781 log2 Fold change. These results may be anticipated in light of the narrow latitudinal range that region II comprises and the low levels of genetic variation previously observed.

Differentially expressed genes between longitudinal extremes within region: Region I

Visualization failed to demonstrate any apparent clustering pattern regarding longitude within region I (Fig. 5). It is however possible to distinguish a slight pattern in the PCA-plot along PC2 in which three eastern, 2 intermediate and 3 western samples form evenly distributed and defined clusters with respect to longitude (Fig. 5B). The three eastern samples, which comprise samples VLA.3, IRRU.2 and BEL.5, range from 104° to 131°. The intermediate samples, SE.14 and SE.33 were collected at 18° and 14° respectively. STJ.2, STA.4, FRA.50, which constitute the western samples, range from -3° to 7°. Of the four samples that interrupt the pattern WAC.5, the western recluse to the right in the plot was collected at - 97°. To certain extent, the samples' clustering patterns thus reflect their longitudinal distribution.



A. B.

Figure 5. Heatmap over all samples in region I with name of individual sample and longitudinal definitions specified (A). PCA-plot over all samples from region I contrasting longitudinal definitions east versus west (B).

As could be expected from the clustering, binomial testing provided merely four significant genes from contrasting longitude within region I, none of which could be included in a cluster when using clustering annotation analysis in DAVID (Table 5). Their functions could therefore not be specified through DAVID, but through Arabidopsis.org. One of the four was up regulated while the remaining three were down regulated (Table 5). This could indicate variation in gene expression in western versus eastern samples. The number of genes are however inadequate for such an estimate to be made. The results from the contrast failed to prove any greater difference or pattern in gene expression between eastern and western samples included in region I.

Table 5. Homologous genes in *A. thaliana* differentially expressed when contrasting longitudinal extremes within region I and II. Locus tags, gene names and functions are specified, as well at their regulation.

East vs west in region I					East vs west in region II				
Locus tag	**Gene Name**	**Function**	**Log2**	**Regulate**	**Locus tag**	**Gene Name**	**Function**	**Log 2**	**Regulate**

			Fold change	d in west relative east				Fold change	d in west relative east.
AT3G54940	AT3G54940	Proteolys, peptidase activity	-2.284	Down	AT2G40100	Chlorophyll a-b binding protein CP29.3, chloroplastic	radiation, lightstimulus	-0.780	Down
AT4G14390	AT4G14390	Unknown	5.991	Up	AT4G31500	Cytochrome P450 83B1	radiation, lightstimulus	2.139	Up
AT5G01670	AT5G01670	*Various	-2.553	Down	AT2G26150	Heat stress transcription factor A-2	radiation, lightstimulus light response and intensity	-2.241	Down
AT5G54610	ANK/ ANKYRIN/ BDA1/ BIAN DA 2	*Various	-3.780	Down	AT4G04020	Probable plastid-lipid-associated protein 1, chloroplastic	radiation, lightstimulus light response and intensity	1.505	Down
					AT1G54050	AT1G54050	radiation, lightstimulus light response and intensity	-2.263	Down

* Functions such as MAPK cascade, defense response to bacterium, defense response to fungus, innate immune response, jasmonic acid mediated signaling pathway, negative regulation of defense response, negative regulation of programmed cell death, plasma membrane, protein targeting to membrane, regulation of hydrogen peroxide metabolic process, regulation of innate immune response, regulation of plant-type hypersensitive response, response to salicylic acid stimulus, salicylic acid biosynthetic process, salicylic acid mediated signaling pathway, systemic acquired resistance, systemic acquired resistance, salicylic acid mediated signaling pathway .

Differentially expressed genes between longitudinal extremes within region: Region II

The pattern indicated by clustering analysis of all samples in region II, excluding sample HRB, is scattered with respect to longitude (Fig. 6).

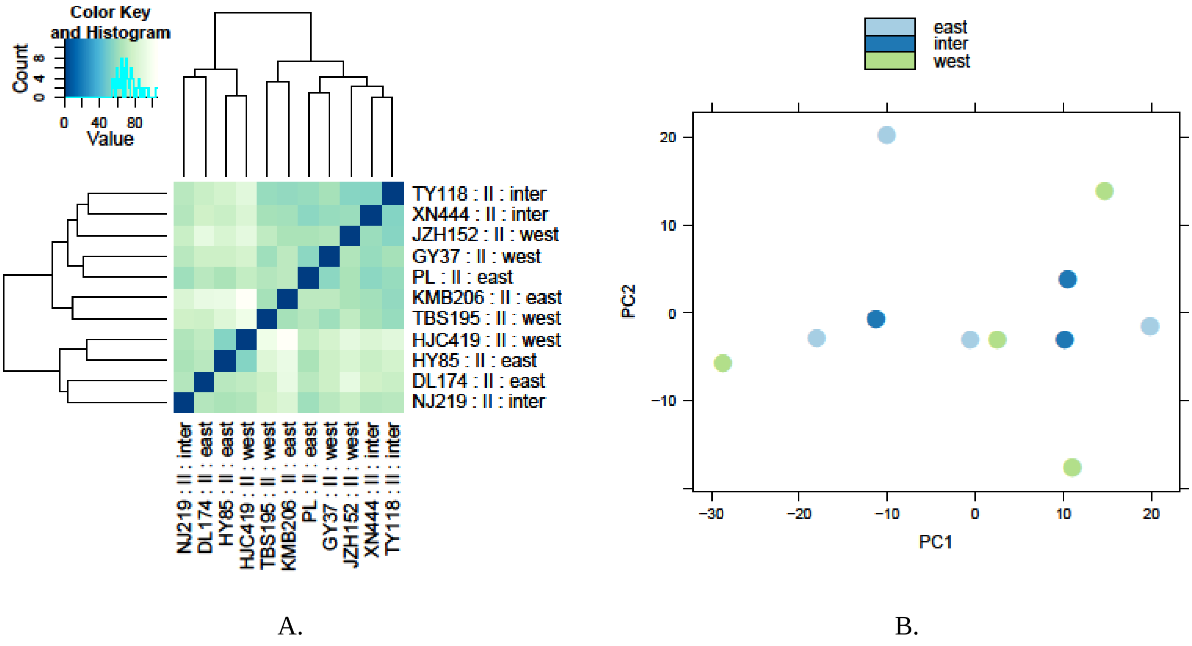


Figure 6. Heatmap over all samples in region II excluding sample HRB, with name of individual sample and longitudinal definition specified (A). PCA-plot over all samples from region II with HRB sample omitted, contrasting longitudinal definitions east versus west (B).

94 differentially expressed genes were presented when contrasting western and eastern extremes in Region II, excluding intermediate values, as well as HRB. The five clusters with the highest enrichment scores ranged from 1.85 to 1.18 and were thus lower than the scores generated for clusters when contrasting region I and II, as well as latitude within region I (Table 6). The clusters were related to various biological functions (Table 6). The difference in gene expression observed when contrasting longitude in region II is best explained by demographic events, as no apparent patterns in biological function are evident.

However, the last cluster contained 5 the genes involved in light detection and response (Table 6). All but one were down regulated. Although this dissimilarity in regulation may indicate differences in gene expression the amount of genes is inadequate for suggestions to be made.

Table 6. Results from DAVID representing the five clusters with the highest enrichment scores in gene lists generated from contrasting longitude in region II.

Term	Count*

Annotation cluster 1 Enrichment Score: 1.85	
response to heat	5
stress response	5
response to temperature stimulus	7
cytoplasm	7
response to inorganic substance	7
phosphoprotein	7
response to abiotic stimulus	10
Annotation cluster 2 Enrichment Score: 1.71	
Plant lipid transfer protein/seed storage/trypsin-alpha amylase inhibitor	4
AI	4
lipid binding	5
lipid transport	4
Plant lipid transfer protein and hydrophobic protein, helical	3
lipid localization	4
Annotation cluster 3 Enrichment Score: 1.43	
response to hydrogen peroxide	4
response to reactive oxygen species	4
response to inorganic substance	7
response to oxidative stress	5
cellular response to stress	5
Annotation cluster 4 Enrichment Score: 1.19	
cell wall	9
external encapsulating structure	9
signal	11
Secreted	8
glycoprotein	8
apoplast	5
extracellular region	10
plant-type cell wall	4
signal peptide	11
disulfide bond	5
glycosylation site:N-linked (GlcNAc...)	7
disulfide bond	5
Annotation cluster 5 Enrichment Score: 1.18	
response to high light intensity	3
response to light intensity	3
response to light stimulus	5
response to radiation	5

*Count values presented in the table designate the number of genes included in the corresponding term. The same genes may occur in several clusters and terms and hence be counted several times.

Verification with baySeq

The differentially expressed genes provided by contrasting region I with II over all samples in DESeq were verified with baySeq (Hardcastle, 2009). The latter generated 621 differentially expressed genes when contrasting the two regions over all samples, while results from DESeq called upon 2,328 significant genes when performing the same contrast. Gene lists in DAVID presented five genes associated with circadian rhythm and rhythmic processes, five with the detection of red and/or far-red light, and 13 with registering light stimulus and/or radiation (Table 4). In comparison, the results from DESeq presented 9 genes related to circadian rhythm and 198 genes related to response to light as well as the distinction between red and far-red light (Table 1). Other than this, the test in DESeq also included 6 related to photoperiodism, 6 to photo reduction (Table 1). There is a significant difference in the amount of genes presented by the two methods, in which DESeq displays far more genes than baySeq. This suggests the Bayesian methods used in baySeq to be more conservative, possibly due to that DESeq fails to take into consideration any posterior underlying patterns.

Discussion

The results from this study show that the variation in patterns of gene expression between regions is greater than the variation within each region and that the difference between regions is reflected in variations in patterns of gene expression.

The amount of differentially expressed genes is significantly greater when regions are contrasted than in tests in which latitude and longitude are contrasted within each region respectively. This could be explained by differences in evolutionary history. It has previously been suggested, although not confirmed, that *C. bursa-pastoris* and *C. rubella* have evolved under partial sympatry in the region I, thus enabling introgressive hybridization. In region II, however, only *C. bursa-pastoris* may be found and introgression thus absent (Slotte *et al.* 2008). Further differences previously found between regions I and II concern genetic variation. Region II namely presents a lower degree of genetic variation than region I, suggesting that it is a younger and smaller population, perhaps even a subpopulation originated from region I (Slotte *et al.* 2008, Holm 2010). The differences between regions may thus reflect the different evolutionary histories of the populations.

Differences between regions may also be observed in the variations in patterns of gene expression presented when contrasts were performed within each separate region. Region I exhibited more significantly differentially expressed genes than region II when latitude was contrasted within each region (620 versus 3 differentially expressed genes respectively). The reverse relationship was however presented in contrasts concerning longitude. In this case region I presented fewer differentially expressed genes than region II (4 versus 94 differentially expressed genes respectively). This is in accordance with previous studies of the same data set, which have proven significant differences in gene expression between the two regions. One such study discovered a positive correlation between period length and days to flowering in region I, whereas no such association could be identified in region II (Holm 2010). Thus, patterns of clinal variation in flowering time and in circadian period length were not equally prominent in the two regions (Holm 2010). The results from the contrasts within each region not only supplement the observed differences between regions, but may also aid in revealing what some of these differences may be due to. Correlations of gene expression with latitudinal clines most likely reflect local adaptations (Mitchell-Olds and Schmitt 2006). This correlation is more prominent within region I, which indicates different local adaptations in the two regions. Correlations with a longitudinal cline may be due to random processes caused by demographic history, as might be the case within both regions, but in particular perhaps within region II. Again, differences in patterns of gene expression are shown to reflect the different evolutionary histories of the populations. These differences may be adaptive or non-adaptive.

There are however differences between the two regions, besides local adaptation and demographic history, that may affect the results. In particular, the two regions cover different latitudinal ranges. Region II presents a comparatively narrow latitudinal range, which may decrease the probability of detecting significant variation due to local adaptation or demography within the region. It also comprises a larger variation in local conditions, which may reduce this probability further. As mentioned above, this region has been previously noted to contain less genetic variation than region I (Slotte *et al.* 2008, Holm 2010). This may decrease the possibility of distinguishing differences within region II, which would also affect the contrast between the two regions.

Studying the widespread coniferous weed *C. bursa-pastoris* aids in answering questions concerning evolution and ecology. How is the rapid and wide expansion of the species reflected in patterns of variation in gene expression? How do local adaptations to a wide variety of environments affect levels of gene expression? This study has observed differences in gene expression within the species putatively caused by such evolutionary events due to the regional division of worldwide sample of *C. bursa-pastoris*. It would, however, be interesting to further investigate the differences between the regions as shown by the amount and sort of differentially expressed genes. Do the patterns of a specific genes' expression reflect the regional division? This could be achieved by comparing a specific genes' expression level in different contrasts. Further understanding concerning the differences due to the division of the dataset into the two regions may also be obtained by testing different sets of contrasts. One could redefine the latitudinal/longitudinal/intermediate groupings to see in what way this would alter the results. This could *e.g.* show to what extent the observed differences depend on the difference in clinal range exhibited by the two regions.

Although speculations concerning the nature of the differences between the regions can be made *e.g.* local adaption and demographic history, this study is insufficient in establishing what evolutionary mechanisms that may have caused the observed differences, as these have not been tested. What evolutionary factors have driven the divergence of these two regions? This could be answered by studying whether expression levels of a particular gene correlate to the ecological importance of that gene, *e.g.* by comparing a specific genes' expression level in different contrasts. One could also study in what way a specific genes' expression is altered by running different contrasts and whether the expression levels of the genes presented from contrasts reflect their position as defined in the contrast, *i.e.* if genes with intermediate gene expression also represent intermediate latitudinal or longitudinal values. Furthermore, modifying the regional division as to better obtain well-characterized environments, as well as increasing the sampling, could better shed light on the evolutionary instances that have taken place.

This study has focused on patterns of gene expression as an intermediate phenotype. Further studies regarding the accuracy of such tests would be fruitful. Are the different levels of gene expression correlated to nucleotide variation or the promoter region of the gene? Do these genetic variations in turn correlate to clinal patterns of variation? And, as to better differentiate between adaptive and random changes; are these variations correlated to fitness? Studies set out to answer these questions aid in increasing our understanding of evolutionary mechanisms and processes.

Acknowledgements

I would like to express my gratitude to the two people who made this project possible. Sincerest thanks to my supervisor Karl Holm for the inspiration, encouragement and enthusiasm in his guidance. I would also like to express appreciation to my good friend and fellow student Johanna Nyström for always knowing when a push, shove or hug is needed.

References

Bioconductor (2004): Open software development for computational biology and bioinformatics R. Gentleman, V. J. Carey, D. M. Bates, B.Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, and others 2004, Genome Biology, Vol. 5, R80.

Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA. 2003. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biology 4:R60.

Hardcastle T J. 2009. baySeq: Empirical Bayesian analysis of patterns of differential expression in count data. R package version 1.12.0.

Hardcastle T J, Kelly K A. 2010. baySeq: Empirical Bayesian analysis of patterns of differential expression in count data. BMC Bioinformatics 11:422

Holm K, Gould P D, Hall A, Lascoux M, Lagercrantz U. 2010. Natural variation in circadian rhythm in a worldwide sample of *Capsella bursa-pastoris* (Brassicaceae). Studies on natural variation and evolution of photoperiodism in plants. Manuscript. Uppsala University.

Holm K, Peele H, Lascoux M, Lagercrantz U. 2010. Genetic basis for correlated variation in circadian rhytm and flowering time in tetraploid weed *Capsella bursa-pastoris* (Brassicaceae). Manuscript. Uppsala University.

Huang H, Yan P, Lascoux M, Ge X. 2012. Flowering time and trasncriptome variation in *Capsella bursa-pastoris* (Brassicaceae). New Physiologist 194:676-689.

Hurka H, Friesen N, German D A, Franzke A, Neuffer B. 2012. Molecular ecology 21:1223-1238.

Mitchell-Olds T, Schmitt J. 2006. Genetic mechanism and evolutionary significance of natural variation in *Arabidopsis.* Nature 441:947-952.

R Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Simon A, Wolfgang H. 2010. Differential expression analysis for sequence count data. Genome Biology 11: R10.

Strickler S R, Bombarely A, Mueller L A. 2012. Designing a transcriptome next-generation sequencing project for a nonmodel plant species. American journal of Botany 99(2):257-266.

Supplementary data

As a means to assess the quality of the data set, variance estimation was performed for each test, in which the dispersions around a gene are estimated. The graph shows a comparison of dispersions and shot noise, both of which are considered in differential expression inferences. The latter is dominant for genes with low expression while the former is prominent for highly expressed genes. The data below the regression line is defined as sampling variance and thus shifted upwards to the values predicted by the line. The data above the line represents the "true” variance and is hence not adjusted. The ratio of gene dispersion and shot noise indicates how well the data accord to the expectations of the experiment.

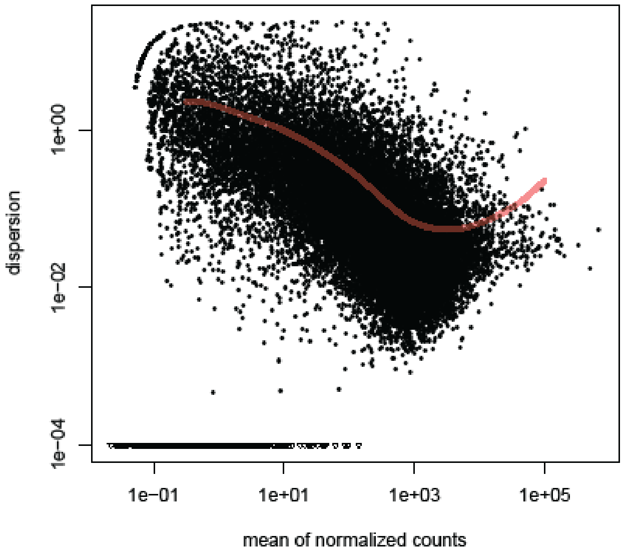
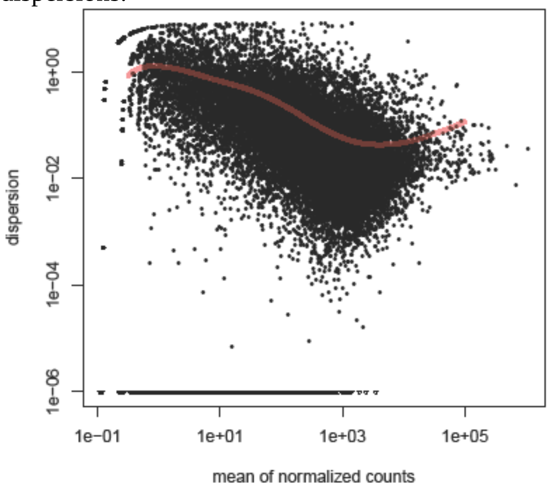
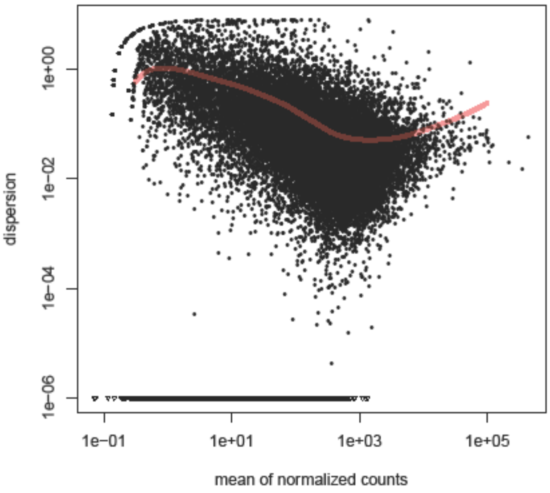


Figure S1. Dispersion values plotted against the mean of the normalized counts when contrasting region I with II and omitting sample HRB. The black dots comprise the empirical and the red line the fitted dispersions.



A



B

Figure S2. Dispersion values plotted against the mean of the normalized counts when contrasting northern and southern extremes and omitting intermediate samples within region I (A).Dispersion values plotted against the mean of the normalized counts when contrasting northern and southern extremes and omitting intermediate samples within region II (B). See figure S1 for further details.

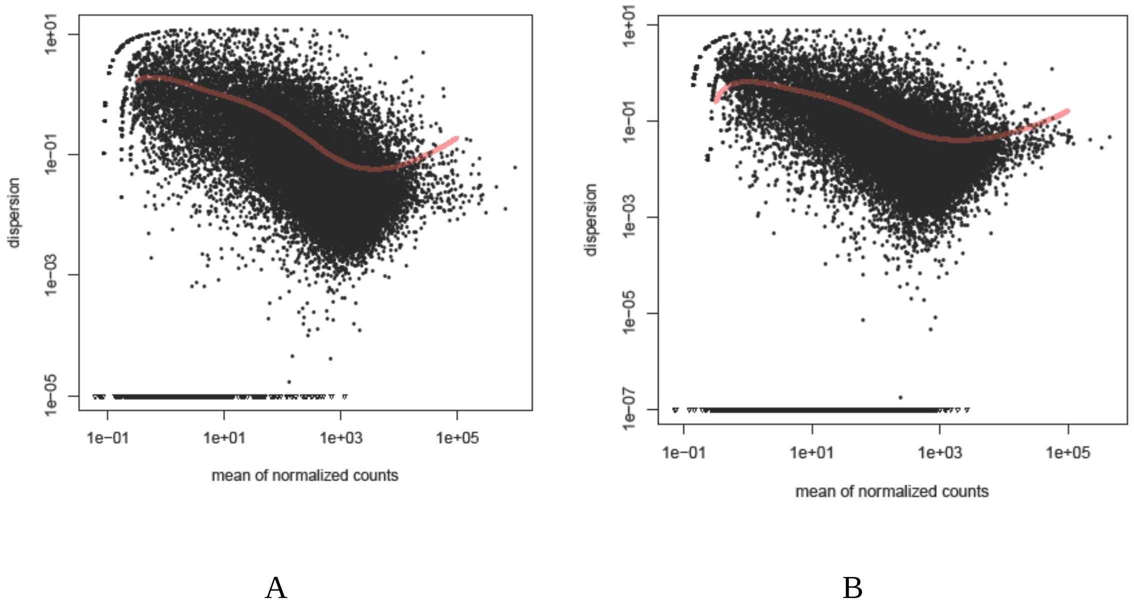


Figure S3. Dispersion values plotted against the mean of the normalized counts when contrasting eastern and western extremes and omitting intermediate samples within region I (A). Dispersion values plotted against the mean of the normalized counts when contrasting eastern and western extremes and omitting intermediate samples within region II (B). See figure S1 for further details.