Molecular Methods in Genome Research

Introduction

Arabidopsis thaliana is the model organism for plant science. The small flowering plant, which is a member of the family Brassicaceae, has many beneficial properties. Diploid genetics with a small genome size and relatively low repetitive DNA content, but also its short life cycle of six weeks with a large seed output recommends the plant as a model for genomic research (Goodman et al. 1995).

Although *A. thaliana* is native to Europe and Asia, it is naturalized across the world with wide over 1,000 known accessions (Horton et al. 2005). Because the natural habitats span a wide geographic range and a variety of climate conditions, there are local adaptations and substantial differences between the accessions making them interesting for studies. When the *A. thaliana* genome sequencing of the accession Columbia-0 (Col-0) was completed and published in 2000 as the first plant genome, it was the corner stone for all further research. For example, research fields deal with the distribution of transposable elements (TE), gene duplication and the comparative analysis of Arabidopsis accessions (The Arabidopsis Genome Initiative 2000). TEs are discrete DNA segments able to move and replicate within genomes. They comprise a great part of the middle repetitive DNA of plant genomes and cause a genetic variability, which ranges from changes in the size and arrangement of the whole genome to changes in single nucleotides. These mutagenic effects are not simply produced by the initial insertion, but through their excision, leaving small indications of their former presence behind (Kidwell et al. 1997).

Four different experiments are described here, dealing with the variations of four *A. thaliana* accessions. The first experiment concentrates on TEs. For this purpose numerous primers and PCR approaches were used. First of all, a variation PCR is performed to find differences and similarities. The results enabled the construction of a phylogenetic tree showing the genesis of the TEs in the studied accessions. Also, own primer pairs were generated to check two areas on chromosome 2 of a TE of Col-0 and a long range amplification PCR was done to complement the results of the variation PCR.

The second experiment deals with the examination of the *SEC10* locus. *SEC10* encodes a core subunit of the exocyst tethering complex (Eckardt 2008). Based on the reference sequence, it was first thought to be a single copy gene. But a non-Mendelian segregation in *A. thaliana SEC10* mutants and a re-sequencing pointed to a hidden gene duplication (Vukasinovic et al. 2014). In this experiment, the *SEC10* locus of the four *A. thaliana* accessions should be investigated in order to give a hint if the *SEC10* duplication occurred prior to separation of all accessions. Therefore, a PCR was done to amplify the whole *SEC10* locus in several fragments.

The examination of a 1.2 Mbp inversion on chromosome 4 is the third experiment. It is known that two different alleles exist of this locus (Zapata et al. 2016). However, Nd-1 displays an inversion compared to Col-0 which is different from the two known alleles (Pucker et al. 2018). So the task here is to check via PCR which allele is present in the four accessions.

The last and fourth experiment looks into the centromeric repetitive sequences. A quantitative real-time PCR is performed to specify the number of the 180 bp long repeats (Brandes et al. 1997).

Material and Methods

The start and the basis for all the following experiments was the DNA extraction of four different accession of the model plant *Arabidopsis thaliana*. For this the CTAB method was used and two replicates from each plant were made. The quality of the DNA was checked by a NanoDrop spectrophotometer. The four different plant accessions are named in the table below.

Tab.1: Numbering and names of the used A. thaliana accessions.

Plant ID	Accession	Name
1	N22564	RRS-7
11	N22622	Wei-0
24	N22596	HR-5
25	N22597	HR-10

Accession 1, RRS-7, is native to the USA, more precisely to North Liberty. With 38-42 cm in height it's one of the larger accessions. The phenotype is characterized by large rosettes and leaves with elongated petiols. The leaf margins are almost smooth and the plant flowers late. In comparison, accession 11 Wei-0 from Switzerland, is smaller with 32-37 cm. So the rosette is medium sized and flatten. The leaf margins are almost smooth as well, but Wei-0 has an early flowering. The UK HR-5 accession differs from the other plants. It is the smallest one with 30-35 cm and medium sized rosettes. The leaf margins are serrated. HR-5 is also typified by numerous axillary inflorescences and an intermediate flowering time. The last one, accession 25 HR-10, is also from the UK, but with 47-53 cm in size a lot higher. It has medium sized rosettes and the leaf margins are almost smooth like the first two accessions. But it is characterized by numerous axillary inflorescences and an intermediate flowering time like the HR-5.

Structural Variations

To define structural variations from the four different accession of *Arabidopsis thaliana* several PCRs were run. The aim was to identify insertions and transposable elements.

Variation PCR

The variation PCR was carried out with the GoTaq® DNA polymerase (Promega). For further details see the GoTaq® PCR protocol. The table below shows the primer pairs with the expected product size, which are transposable elements from the Nd-1 or Col-0 accession. The assignment is unknown.

Tab. 2: Primer pairs for the variation PCR with annealing temperature and expected product size.

Primer Pair	fw Primer	rv Primer	Annealing [°C]	Expected Product Size [bp]
a	Nd02	Nd03	61	820
b	Nd08	Nd09	60	1500
c	Nd10	Nd11	61	1000
d	Nd20	Nd21	61	1500
e	Nd30	Nd31	61	1000
f	Nd32	Nd33	61	500
g	Nd36	Nd37	60	1000
h	Nd38	Nd39	61	650

LongAmp PCR

The approach here is similar to the variation PCR. The aim is to identify a 13kb long fragment. The first three PCRs are carried out with a GoTaq® DNA polymerase (Promega), which is unable to amplify such a fragment. But different primer pairs were used (see Fig. 2), which are binding within the putative insertion. The expected product size of the primer pairs are much shorter and serve as a comparison for the LongAmp® Taq (NEB) PCR.

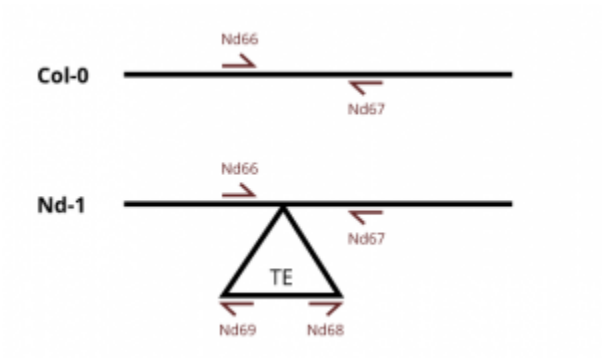

Fig. 1: Primer arrangement for the LongAmp® PCR. With the primer pairs Nd66/Nd69 and Nd68/Nd67 it is only possible to generate an amplicon on the sequence of the Nd-1 allele.

Tab. 3: Primer pairs for the GoTaq® PCR with annealing temperature and expected product size.

Primer Pair	fw Primer	rv Primer	Annealing [°C]	Expected Product Size [bp]
i	Nd66	Nd67	58	300
j	Nd66	Nd69	58	500
k	Nd68	Nd67	58	500

The LongAmp® Taq (NEB) is able to amplify up to 30 kb and is needed here for a 13 kb fragment. For the detection the same primer pair (i) was used as in the previous approach to ensure consistency between experiments. The approach was done based on the LongAmp® PCR protocol.

Tab. 4: Primer pairs for the LongAmp® Taq PCR with annealing temperature and expected product size.

Primer Pair	fw Primer	rv Primer	Annealing [°C]	Expected Product Size [kb]
i	Nd66	Nd67	58	13

Designed Primers

The last experiment for detection of TEs, are two PCRs with self designed primer pairs. They are based on the reference sequence of Col-0 (TAIR10) and should check the unknown accessions for TEs at two different loci on chromosome 2. The PCR was based on the GoTaq® PCR protocol.

Tab. 5: Self designed primer pairs with sequence and annealing temperature.

Primer ID	Name	Sequence	T_M [°C]
N391	SFS_Loc1_fw	ccattagcctttgtaactaggg	54
N392	SFS_Loc1_rv	accggcgatacctagtac	54
N393	SFS_Loc2_fw	cacttcctttgcctgtgtctg	57
N394	SFS_Loc2_rv	tctttagtgctgacacctgac	56

SEC10 Locus

To check if there are two copies of the *SEC10* gene in the four accessions, the primer pairs shown in Fig. 2 were used following the Q5® PCR protocol. The Q5® High-Fidelity DNA polymerase is a recombinant enzyme which has a proof reading activity and a

higher rate of amplification than a classical Taq polymerase. All five DNA fragments were amplified and their quality was checked by gel electrophoresis. The positive PCR products were cleaned-up following protocol of NucleoSpin® Gel and PCR Clean-up (Macherey & Nagel) and cloned into the pMiniT vector using a PCR cloning kit (NEB). For the duplication of the genetic material a heat shock transformation into competent *E. coli* was performed. The bacterial strain was XL-1 Blue. At the end, the transformed bacteria were grown on LB medium with ampicillin as selection marker, because of the ampicillin resistance gene on the pMiniT vector. If possible, 10 to 20 colonies were picked and these were first re-streaked or liquid cultures were directly inoculated. At the same time a colony PCR after the GoTaq® PCR protocol was performed to check if the *E. coli* contained the vector with the DNA fragment. The primer pair N179/N105 was used to determine the size of the insert. The plasmids from the positive clones were harvested and prepared for the sequencing after the TELT Microquick method (Weisshaar 1998). The samples were submitted for sequencing. In order to analyze the sequencing results, the sequence was search via BLASTn against the genome sequences of Col-0, Nd-1, and L*er*.

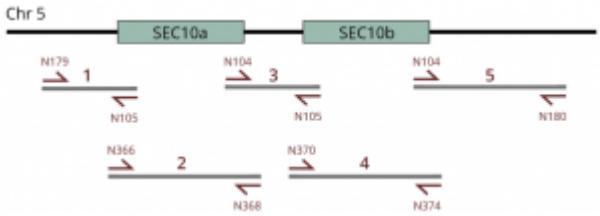

Fig. 2: Primer arrangement for the *SEC10* locus. In total, 5 primer combinations are used. The production of amplicons is considered as evidence of the presence of duplicates

Inversion on Chromosome 4

The ~1.2 Mbp inversion is localized in the north of chromosome 4. The Nd-1 allele is an inversion compared to Col-0. The four unknown accession were investigated to determine if they have the Col-0, L*er*, or Nd-1 allele. The method was a GoTaq® PCR. The primer combination, which was used, can be seen in Fig. 3. The gel electrophoresis should show two bands as the result for every accession to identify either the Nd-1 or the Col-0 allele.

Fig. 3: Primer arrangement for the potential inversion on chromosome 2 with the Nd-1 and Col-0 allele.

Centromeric Repeats

Every *A. thaliana* accession has an individual number of 180bp long repeats around the centromeres. To determine the number of repeats a qPCR is performed. The Luna® Universal qPCR Master Mix (NEB) was used, but the total volume of the reaction was changed to 10µl and the volume of the used master mix per reaction was 5µl. The primer pair for the CHS gene was N377/N378 and the primer pair for the repeats was N375/N376. As reference for the number serves the chalcone synthase gene (CHS), that occurs only once in the genome. Each sample was measured in triplicates and their mean was calculated. Important is the take-off value. This determines the cycle number when the measured fluorescent intensity is higher than the fluorescent background. The deviation between the cycle number of the sample and the CHS reference correlates with the number of the centromeric repeats. During each cycle, the genetic material is doubled.

Results and Discussion

Variation PCR

The results from the structural variation PCR with the following gel electrophoresis (Fig. 4) are summarized in the table below. Presence of PCR products in all lanes indicates that the DNA extraction was successful and the primer pairs were working.

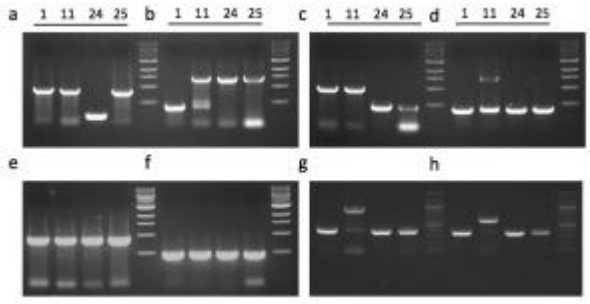

Fig. 4: Gel electrophoresis of the structural variation PCR products. a-h are the used primer pairs. For further information see the table (Tab. 6).

Tab. 6: Measured product sizes in comparison with the expected product sizes in the gel pictures from the variation PCR.

		Measured Product Size [bp]			
Primer Pair	**Expected Product Size**	**Accession 1**	**Accession 11**	**Accession 24**	**Accession 25**

a	820	800	800	300	800
b	1500	400	1400	1400	1400
c	1000	900	900	400	400
d	1500	400	400 & 1500	400	400
e	1000	800	800	800	800
f	500	400	400	400	400
g	1000	350	1000	350	350
h	650	350	600	350	350

In detail, the expected fragment size from each primer pair (a-h) was not observed for all accession. The band sizes measured from accession 1 matched the expected band sizes from primer pairs a,c,e, and f. The rest is shorter than the expected size. It means that the accession 1 has the TEs TE-a, TE-c, TE-e and TE-f but it is missing the rest. The accession 11 has all the TEs, which can be recognized in the agreement with the expected band sizes. The accession 24 has the least matches with the expected product sizes. It only contains the TEs TE-b, TE-e and TE-f. The last accession, accession 25, contains one more TE, TE-a, in comparison with the accession 24.

Examining the similarity of the individual accessions to each other, the great similarity of 24 and 25 can be seen, because they only differ in the TE-a. The next higher similarity shows the accession 1. That has additionally the TE-c, but lost the TE-b. Accession 11 differs most from the other accessions, because it contains at least three TEs more. Based on these results, a phylogenetic tree was constructed to show in which order the TEs probably evolved in the four accessions and how they were related to each other. The phylogenetic tree implicates, that accession 24 first split off first. Afterwards, accession 25 and then accession 1 split off. This result applies only for these eight examined loci. It is a first clue how the four lines are related to each other, but more loci have to be examined to validate these preliminary results.

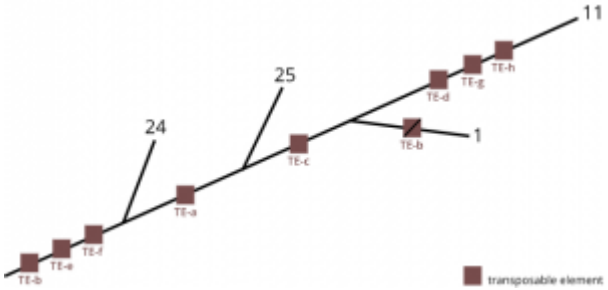

Fig. 5: Phylogenetic tree of the transposable elements (a-h) with the four *A. thaliana* accessions 1, 11, 24 and 25. The greatest similarity share the accession 24 and 25, followed by accession 1. Accession 11 contains all eight examined TEs and thus differs most from the others accessions.

Long Amp® Taq PCR

Two different PCR approaches were applied. Three PCRs with a GoTaq® polymerase for the primer combinations i-k and a PCR with a LongAmp® Taq polymerase for the primer combination i were performed. The gel electrophoresis for the primer combination i shows for the accession 1 the expected product size of 300 bp. The other four accessions lack a band. This result implicates that the accession 1 has the Col-0 allele (see Fig. 1) and that the other accessions could have the Nd-1 allele with the TE. This assumption is supported by the next two gel pictures showing the PCR products for the primer combinations j and k. One band each for accessions 11 and 24 can be seen on the pictures for these primer combinations, which are binding within the TE. If the accession would lack the TE, there should not be a band in these pictures. Because accession 1 did not show a band in any of the three pictures, the result is not evaluable. Possible causes are mistakes in the PCR approach or a lack of DNA quality.

According to these results, a band for the LongAmp® PCR products should be seen for the accessions 11 and 24, but this is only true for accession 24. This observation indicates that the accession 24 contains the 13 kb insertion. The accession 11 has the insertion probably as well, but the quality of the DNA sample might be too low, so that the polymerase is not able to amplify the whole fragment.

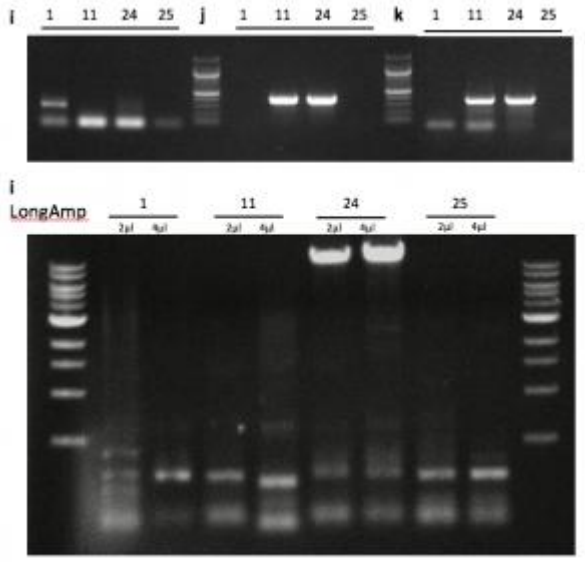

Fig. 6: Gel electrophoresis of the structural variation PCR. i-k are the used primer pairs. For further information see the table in methods. A LongAmp® PCR approach for primer pair i.

Own Primer Pairs

Both primer pairs were designed based on the Col-0 reference sequence. The first four samples in the gel electrophoresis picture (Fig. 7) are the PCR products from the four accessions for locus 1. While accessions 1, 11 and 25 show a band of 300 bp, the accession 24 shows a band of 1300 bp. It means that the accession 24 possesses the Col-0 allele with the TE insertion on chromosome 2, while the other three accessions are lacking the TE. The second part of the picture does not reveal any primer bands at all.

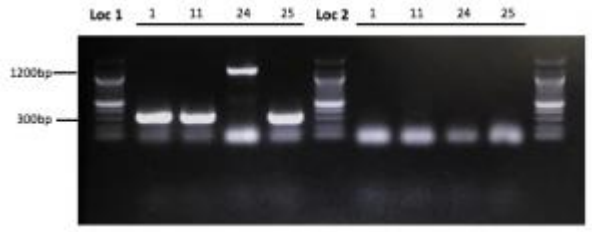

Fig. 7: Gel electrophoresis of the PCR products from the own primer pairs.

Another PCR was performed to assess lower annealing temperatures (data not shown). This was not successful. It could be possible that each of the plants lack the primer binding sites, because the DNA sequences differ too much from the Col-0 sequence. Issues with the primer pairs itself could be another explanation. An error during the synthesis or a wrong sequence could be possible, as well as the formation of a secondary structure.

***SEC10* Locus**

First, the five *SEC10* fragments were amplified with a Q5® PCR for all four accessions and validated via gel electrophoresis. All accessions showed a band of the right size except for accession 11 when checked with the fifth primer combination (data not shown). This sample was excluded from further analyses. The other PCR products were cleaned-up, cloned into pMiniT and transformed into chemocompetent *E. coli*. There was a very variable growth of the *E. coli* after one day in the incubator. Some plates showed no colonies, some only a few and some a bacterial lawn. So it was not possible to continue the analysis of all accessions. Following, the focus was on accession 1. The colony PCR revealed, that the bacteria did not contain all the five *SEC10* fragments, but only fragment 1 and 3 (Fig. 8). This result is not surprising, because the other three fragments have a size of 7 kb. So they are much larger than the other two fragments and this indicates that the chance is lower that they are cloned into the vector. Due to the much larger size of the vector, bacteria are less likely to take it up. The growth of colonies does not have to say, that the colonies all contain the pMiniT with the right

insert. It is likely that they contain either an empty vector or nothing. *E. coli* cells are able to tolerate low amounts of ampicillin and it has a relatively short half-life time. Therefore, growth of false positives on the plates cannot be excluded.

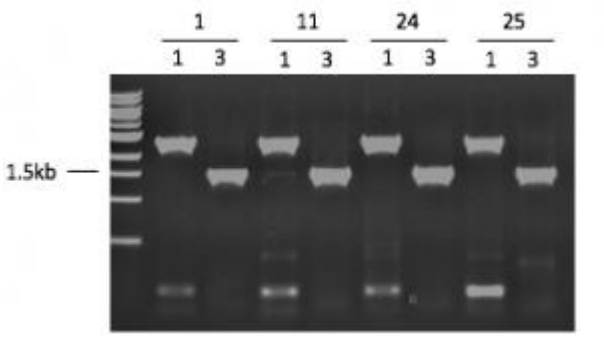

Fig. 8: Gel electrophoresis of the Q5® PCR products of all accessions for the SEC10 fragments 1 and 3. All PCR products were tested positive.

The vectors with the fragments 1 and 3 for accession 1 were harvested from the *E. coli* liquid cultures and sent for sequencing. It has been revealed that the sample with the first fragments has not got the necessary DNA content and has not been sequenced. The actual size of the insert (fragment 3) can be seen in Fig. 9. The gel electrophoresis picture was taken right after the Q5® PCR and it shows a size of 1.5 kb for the third fragment for each of the four accessions.

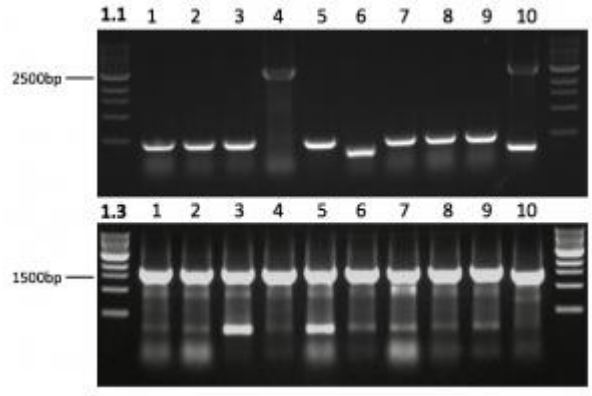

Fig. 9: Gel electrophoresis of the positive tested colony PCRs. Shown is the gel from accession 1 fragment 1 and accession 1 fragment 3 with 10 tested colonies each.

The table below (Tab. 7) displays the results from the sequencing. Therefore, the obtained sequence from the forward and the reverse primer was mapped against the genome sequences of Col-0, Nd-1 and L*er* individually. This data revealed the beginning and the end of the *SEC10* fragment on the chromosome of the reference genes and thus the actual length could be calculated. The results for L*er* and Nd-1 are consistent with the fragment size of 1.5 kb. Also the identity is for all samples around 99 % which indicates a correct sequence, because Sanger sequencing has an error rate of 1 %. However, Nd-1 has three base pairs more. It cannot be said if the Nd-1 accession has an insertion or the L*er* accession a deletion. But this result is validates the hypothesis that the accession 1 has two copies of *SEC10* in the locus. Otherwise, the fragment 3 would not have been amplified. Furthermore, there has been a complete different result for Col-0. When the Sanger sequencing read was search via BLASTn against TAIR10 it resulted in a fragment size of -5748 bp. This can be explained by the collapse of this region in the Col-0 reference sequence. The low alignment identity of 86-88 % underlines this. It could be possible, that Col-0 has one matching sequence upstream of the *SEC10a* region and the other one downstream of it (Fig. 2). This would explain the much bigger size of the fragment.

Tab. 7: Results of the sequencing and blasting from the SEC10 fragment 3.

	Primer	**Location**	**Identity [%]**	**Sequencing Start**	**Sequencing End**	**Sequenced Length [bp]**	**Fragment Size [bp]**
Col-0	NEBf	Chr 5	88,21	4008419	4008951	532	-5748
Col-0	NEBr	Chr 5	86,83	4002671	4002355	316	
Nd-1	NEBf	NdChr 5	99,28	3984459	3985290	831	1553
Nd-1	NEBr	NdChr 5	98,31	3986012	3985132	880	
L*er*	NEBf	gi|1032275144	99,64	4048529	4049360	831	1556
L*er*	NEBr	gi|1032275144	99,32	4050085	4049202	883	

Inversion on Chromosome 4

In order to detect if the 4 accessions contain the Col-0 or the Nd-1 allele, four different primer pairs have been used in a PCR. Two pairs each amplify the beginning and the end of the inversion region (Fig. 4). It can be seen in the gel electrophoresis of the PCR products, that the accession 1, 11 and 24 contain the Nd-1 allele for the left breakpoint, but probably the L*er* allele for the right one. Accession 25 only shows one band, the Col-0 allele for the right breakpoint. So it cannot be said if accession 25 contains the Col-0 allele or mixed alleles as the other four accession. Accession 25 has probably not the right binding site for the left breakpoint, so there could not be a PCR product. The most likely explanation is a failure in the PCR. The results of the mixed alleles could possibly be caused by an error in the PCR. Maybe the primers for the breakpoints were switched or the marked positions of the primer pairs on the gel pictures are wrong. But this can only be found out by repeating the experiment.

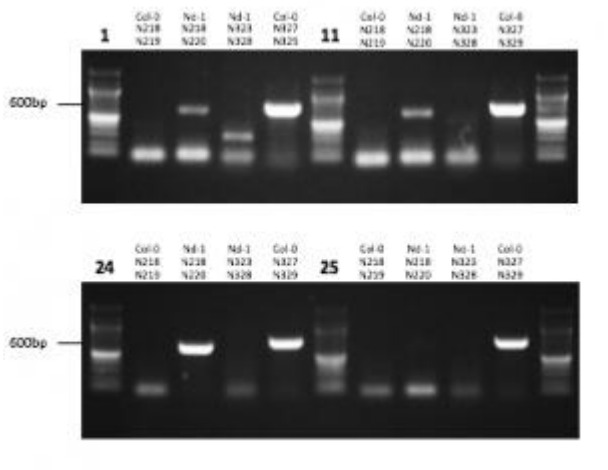

Fig. 10: Gel electrophoresis of the inversion PCR products. Each accession (1, 11, 24 & 25) should have to bands to determine the Col-0 or the Nd-1 allele. But nearly every accession has one band for both alleles.

Centromeric Repeats

The aim of the real time qPCR was to detect the copy number of a 180 bp repeat. The chalcone synthase was used as a reference gene, because of its single copy in the genome. The DNA samples of each accession were measured in triplets and the mean of the Cq value was calculated (Tab.8). Four samples showed a failure and no Cq value was available in the results. But by looking on the raw data (accession 24 replicate 3, accession 25 replicate 2 primer pair repeats and two negative controls for the CHS), a fluorescent signal could be seen but with a much higher starting value. It could be possible that the device was not able to detect the Cq value, because the fluorescent intensity was too high since the first cycle caused by a too high DNA concentration. Dilutions could be a solution for this issue. Especially the samples from accession 25 showed a saturation of the signal after a few cycles. Compared to the other replicates, its values were much higher and could be caused by bubbles in the samples. Where a reliable Cq value was missing for a sample, the mean was calculated by the values from the remaining two replicates. Fig. 11 shows that there are no outliers in the triplicates with exception of accession 1. Therefore, the values of the triplicates were close enough together to keep all values for downstream analyses. But it has to be noted that the number of repeats in the first accession can actually be higher than it is calculated here.

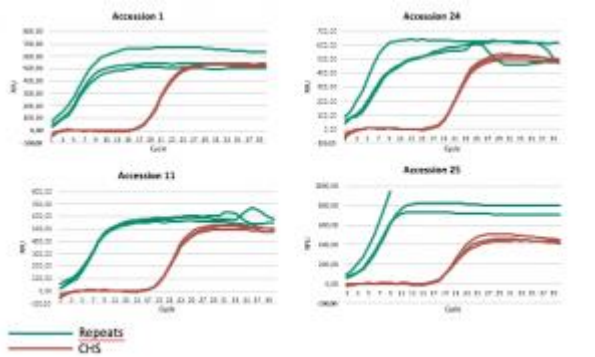

Fig. 11: qPCR curves. Each diagram shows the measured fluorescent of the repeats PCR products (green) and of the CHS PCR products (red).

Tab. 8: Mean of the measured Cq values from the real time PCR.

Primer Target	fw primer	rv primer	Mean Cq Accession 1	Mean Cq Accession 11	Mean Cq Accession 24	Mean Cq Accession 25	Mean Negative Controll
Repeats	N375	N376	2,22	2,95	2,54	2,08	25,20
CHS	N377	N378	18,63	19,40	18,95	19,30	34,90

The first thing to be seen in the table is, that all the Cq values for the examination of the repeats are much smaller that these for the reference gene CHS. This implicates that the qPCR has worked and that all accessions have more than one copy of the 180 bp repeat in its genome. For the calculation of the number of the repeats, the Cq values from sample for the CHS were subtracted from the Cq values from the samples for the repeats for each accession. This result is the difference in the cycle number. Theoretically, the DNA amount is double in each cycle ant thus the fluorescence intensity doubles as well. Therefore, the cycle difference indicates the number of copies of the sequence of interest. The number can be calculated with this formula: $2^{\Delta Cq}$. The results from this calculation are shown in Tab.9.

Tab. 9: Calculated number of the 180 bp repeats

	Accession 1	Accession 11	Accesssion 24	Accession 25
Number of Repeats	86639,00	89677,53	87394,97	152664,00

Except for the result of accession 25, the other accessions contain between 86,000 and 90,000 repeat copies. The accession 25 contains 152,664 copies and that is nearly twice as much as the others. However, sequence divergence from the reference sequence might influence this result. Small Cq value differences can have a huge impact. In this case, the Cq value for the repeats is a little bit smaller than the others and the Cq value for the CHS is a little higher than the values for the other accessions. Therefore, it is possible that the real value is around 88,000 as well.

References

Brandes, A., Thompson, H., Dean, C. (1997): Multiple repetitive DNA sequences in the paracentromeric regions of Arabidopsis thaliana L. Chromosome Research, 5, 238-246

Eckardt, N. A. (2008): An Exocyst Vesicle Tethering Complex in Plants. The plant cell, doi: https://doi.org/10.1105/tpc.108.20051

Goodman, H.M., Ecker, J.R. & Dean, C. (1995): The Genome of Arabidopsis thaliana. Proceedings of the National Academy of Sciences of the United States of America, 92(24), 10831-10835

Horton, M.W., Hancock, A.M. & Huang, Y.S. (2005): Genome-wide patterns of genetic variation in worldwide Arabidopsis thaliana accessions from the RegMap panel. Nature genetics, 44, 212-216

Kidwell, M.G. & Lisch, D. (1997): Transposable elements as sources of variation in animals and?plants. Proceedings of the National Academy of Sciences of the United States of America, 94(15), 7704-7711.

Pucker, B., Holtgraewe, D., Stadermann, K.B. (2018): A Chromosome-level Sequence Assembly Reveals the Structure of the Arabidopsis thaliana Nd-1 Genome and it's Gene Set. bioRxiv, doi: https://doi.org/10.1101/407627

The Arabidopsis Genome Initiative (2000): Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature, 408, 796-815.

Vukasinovic, N., Cvrckova, F., Elias, M., Cole, R., Fowler, J. E., Zarsky, V. & Synek, L. (2014): Dissecting a Hidden Gene Duplication: The Arabidopsis thaliana SEC10 Locus. PLoS ONE, 9(4), e94077. http://doi.org/10.1371/journal.pone.0094077.

Zapata, L., Ding, J., Willing, E.-M. (2016): Chromosome-level assembly of *Arabidopsis thaliana* L*er* reveals the extent of translocation and inversion polymorphisms. PNAS, 113(28), 4052-4060