Norwegian University of Life Sciences

Master's Thesis 2018 60 stp

The Faculty of Biosciences

Åshild Ergon

Identification of red clover (*Trifolium pratense L.)* SNPs whose allelic versions appear with different frequency in pure stands and in mixtures with grasses, using GBS and CAPS-markers

Vegard Eriksen Sæther

Biotechnology - Genetics

Faculty of Chemistry, Biotechnology and Food Science (KBM)

Acknowledgements

There are many people who should be acknowledged for supporting me while I wrote this master thesis. The reasons are many, economical, academical, emotional and many more. First in line is my parents, my mother Gunn Oddny Eriksen and my father Leif Inge Sæther. They have both supported my studies economically from start when I started out with my bachelor-degree in biology and until now when this thesis was done. They also provided encouragement, and belief in my academical progress. I also want to acknowledge those who have supported me academically. The first in line is Åshild Ergon, my supervisor for this thesis. We have had many discussions about theory and writing of this thesis. A big help was also Anne Guri Marøy and Sylvia Sagen Johnsen who both helped me a lot during my days in the laboratory. The last people I want to acknowledge is my friends, both the old ones and the new ones. My old friends lived in other parts of Norway during my studies, but they never forgot about me, and we talked from time to time. I also want to thank my new friends, who helped make my years at NMBU (Norwegian University of Life Sciences) a positive experience.

Ås, May 2018

Vegard Eriksen Sæther

Sammendrag

I denne oppgaven vil jeg prøve å finne ut om det er en genetisk forskjell mellom rødkløver (*Trifolium pratense L.)*-populasjoner i rene bestander med bare rødkløver, og blandede bestander hvor rødkløver har vokst sammen med gressarter. Dette ble gjort ved å studere SNPer som vi fant i rødkløver-genomet. Bladprøver samlet inn i et tidligere eksperiment ble brukt. Bladprøvene ble samlet som prøver med 100 individer hver, fire prøver fra rene bestander og fire prøver fra blandingsbestander. Fra to av ren-bestandene ble det tatt tre replikate prøver. De replikate prøvene ble samlet for å studere mengden med tilfeldig variasjon mellom prøvene. Dette resulterte i tolv prøver totalt som ble sendt til BGI i Kina for genotyping-by-sequencing (GBS) for å finne SNPer. Etter at BGI gjorde en serie filtreringer så mottok jeg et datasett med 129 661 SNPer. Hver av disse SNPene ble oppgitt med hvor mange ganger (antall reads) hver av de to alleliske formene av SNPen forekom i hver av prøvene. For å gjøre videre resultater mer nøyaktige så utførte jeg et par filtreringer av dataene. Jeg fjernet alle SNPer fra datasettet som hadde et summert antall reads utenfor intervallet 100-499 i en eller flere av de tolv prøvene. Jeg fjernet også alle SNPer fra datasettet som hadde en «minor-allele-frequency» under 0.05 i en eller flere av de tolv prøvene. Replikatene som kom fra samme bestand ble så sammenlignet med hverandre, både med utregning av korrelasjon og ved PCA. Det var nesten like stor variasjon mellom replikater fra samme bestand som det var mellom bestandene. I samme PCA var også de andre prøvene med, og den første PCA-aksen forklarte 25% av variasjonen, og delte prøvene inn i ren-bestand og blandingsbestand. For å finne SNPer som hadde alleliske versjoner som forekom med forskjellig frekvens i ren-bestand og blandingsbestand så kalkulerte jeg F_{ST} ved å sammenligne allelfrekvensen i enkelt ren-bestander mot gjennomsnittet av frekvensen i blandingbestandene, og motsatt, for hver SNP. En CHI^2-test basert på F_{ST}-verdiene ble utført for å finne SNPer hvor de alleliske versjonene forekommer med signifikant forskjellig frekvens i ren-bestand og blandingsbestand. CHI^2-testen ble utført på p-nivå 0.1, 0.05 og 0.01. Dette resulterte i 9, 6 og null SNPer, i den rekkefølgen. Jeg forsøkte så å bekrefte resultatene i et annet plantemateriale.

Resultatene som ble brukt videre var de for p-nivå 0.1. Dette ble gjort ved å utvikle CAPS-markører, som betyr at jeg tok i bruk restriksjonsenzymer for å skille forskjellige genotyper fra hverandre. Blader ble samlet inn fra etterkommerne etter populasjonene som ble sendt inn til BGI, samt etterkommerne til populasjoner fra noen andre ren-bestander og blandingsbestander. Forskjellen nå var at DNA ble ekstrahert fra enkeltindivider, ett individ er lik en prøve. Jeg prøvde å utvikle CAPS-markører for fire av SNPene. Primere, og så restriksjonsenzymer, ble testet for sine evner til å skille alleler fra hverandre på en gel, noe som resulterte i at kun to forskjellige SNPer ble genotypet hos enkeltindividprøvene. Genotyping med CAPS-markører viste en annen allelfrekvens enn hva den var i datasettet fra BGI. Det så ikke ut til å være noen forskjell mellom etterkommere fra ren-bestander og etterkommere fra blandingsbestander.

Abstract

In this thesis I`m trying to uncover if there is any genetic difference between red clover (*Trifolium pratense* L.) populations grown in pure stands with only red clover and in mixed stands, where red clover is grown together with grasses. This was done by studying SNPs found in the red clover genome. Leaf samples that were sampled in an earlier experiment were used. Leaf-samples were collected as samples with 100 individuals each, from four different pure stands and four mixed stands. From two of the pure stands we collected three replicate samples. The replicate pool-samples were collected to study the random variation between samples. This resulted in a total of 12 samples which were sent to BGI in China for genotyping-by-sequencing (GBS) to find the SNPs. After BGI did some filtrations, I received a dataset with 129 661 SNPs. In addition, the dataset also included how many times (number of reads) each allelic version of the SNPs appeared in each of the samples. To make further results more accurate I performed a couple of filtrations on the data. I removed all the SNPs from the dataset which had a summed number of reads outside the interval 100-499 in one or more of the twelve samples. I also removed all SNPs from the dataset that had a minor allele frequency below 0.05 in one or more of the samples. The replicates from the same plots were then compared to each other. It

was almost as great variation among the replicate samples from the same plot as it was between plots. The other samples were also analyzed in the same PCA. The first PCA-axis explained 25% of the variation in my samples and divided the samples into a group of pure stands and a group of mixed stands. To find SNPs that had allelic versions that appeared with different frequency in pure stands and mixed stand I calculated F_{ST} by comparing the allele-frequency of single pure stands against the average frequency of the mixed stands and vice versa, and for each SNP. A CHI^2-test based on the F_{ST}-values was performed to find SNPs where the allelic versions appeared with significantly different frequency in pure stands and mixed stands. The CHI^2-test was performed at P-level 0.1, 0.05 and 0.01. This resulted in 9, 6 and zero SNPs, respectively. I chose to try to confirm the results for P-level = 0.1 further. This was done by developing CAPS markers, meaning that I used restriction-enzymes to tell different genotypes apart. Leaves were collected from the descendants of the red clover populations sent to BGI, and the descendant of some other pure stands and mixed stands, but this time DNA was extracted from single individuals. I tried to develop CAPS-markers for four of the SNPs. Primers and then restriction enzymes were tested for their ability to distinguish alleles, resulting in only two different SNPs being genotyped in the individual samples. Genotyping with CAPS markers showed a different allele-frequency than what I got from BGI earlier, and there seemed to be no difference between descendants from pure stands and descendants from mixed stands.

Table of contents

1. Introduction

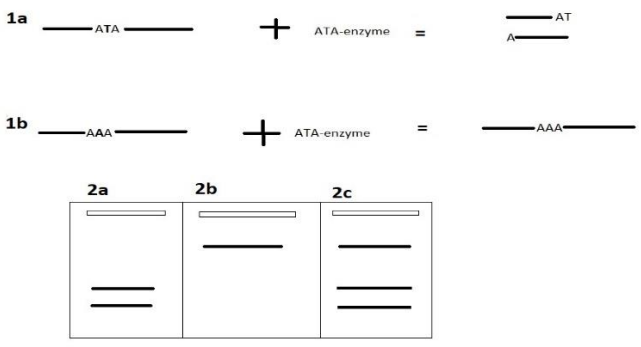
The red clover, *T. pratense* L., is, because of its properties, an important agricultural plant in temperate regions. It perhaps best known for its ability to fix nitrogen from the air with the help of a bacterium called *Rhizobium leguminosarum* biovar *trifolii,* which lives in a symbiotic relationship with the red clovers root. The bacteria fix nitrogen for the red clover to use, and the red clover provides the bacteria with carbohydrates [1]. However, the nitrogen fixed by the bacteria only ends up in the plant itself, so how does this help as a fertilizer? When the rest of the red clover plant somehow is removed, either by grazing, cutting or something else, the roots are left behind. These are still rich with nitrogen, so when they rot the nitrogen ends up in the soil [2]. This can be taken advantage of in agriculture because it offers a more environmentally friendly and cheaper alternative to artificial fertilizers. It`s also a nutritious plant, containing more micronutrients such as minerals and vitamins, and have a higher content of protein than most grass herbage [2]. Having red clover growing in the field together with other kinds of forage also increases the yield. According to different research [3,4,5] a substantial increase in forage yield can be seen in mixtures with red clover and other plants, compared to pure cultures of red clover or other forage plants. This is probably partly a result of the red clovers nitrogen fixation. The red clovers nutritional properties, combined with how well mixed cultures with red clover grow, makes it an excellent plant to grow together with other forage species.

It`s a common problem that red clover isn`t as long lived as the grasses it`s sown among. However, is there any genetic basis in red clover that can be bred upon to improve its survivability? The answer to this can be found by studying the genetics of long-lived survivors of red clover. I want to find areas in the genome that has been under different selection in survivor-populations of pure stands and mixed stands. A genetical difference as such would be different “versions” of areas in the genome which frequencies in the population is affected by selection-pressure provided by the type of culture they are grown in (pure or mixed). Genetic markers are needed to study such areas. There is however many different kind of markers, each with their own

properties. These properties mainly vary in size and frequency they appear along the genome. Some of the genetic markers are long, up to several tens and hundreds of base-pairs long, such as microsatellites. At the other end of the scale there are genetic markers consisting of only a single base/nucleotide, such as single-nucleotide polymorphisms, or SNPs for short. These are quite abundant in the genome of most species. In the human genome SNPs appear with a frequency of 1 every 200th basepair [6]. In maize the same number is about 1 SNP for about every 60 basepairs [7]. SNPs can appear both within a gene and next to it and is therefore very useful as genetic markers, for both animals and plants. How are markers used to identify genes/chromosomal areas? If the genome of the organism has already been sequenced and mapped, a search can be performed with the SNP, with the surrounding sequences, in a database. The search will give the SNPs locus. The search might show that the SNP is inside the gene or a sequence important for its expression. It might also just be closely linked to it. Øystein Milvang [8] in his study where he tries to identify chromosomal areas that control the starting-time for stem elongation in red clover, F_{ST} and CHI^2-test used to find a connection between certain SNPs and timing of stem-elongation. A genetic database was then used to find which genes these SNPs was connected to, and therefore also which genes that might influence elongation timing. The same can be done to find and identify SNPs that appear with different frequencies in survivor populations of pure stands with red clover and mixed stands with red clover. Genotyping-By-Sequencing (GBS) can be used to find SNPs. As described by Elshire [9], GBS is a technique for identifying SNPs by sequencing. One or more genetic libraries are made by digesting the genome with a restriction-enzyme, and then amplify the sequences with PCR. These sequences can then be sequenced by next-generation-sequencing (NGS). However, only the end of the sequences is sequenced, perhaps about 100 bases, leading to what is called “reduced representation of the genome”. This is because we then get a higher number of reads for the parts that are sequenced, making data from these sequences more accurate. Had the whole genome been sequenced we might have gotten more SNPs, but less accurate data. After the end of the sequences have been sequenced, they are aligned. When the sequences have been aligned they can be compared, and the SNPs will reveal themselves as being a single base that is polymorphic among all the sequences.

The SNPs found by GBS is put through a statistical analysis to see which appear with significant different frequencies in pure stands and mixed stands. When SNPs that appear with significant different frequency in these two stands are found, a method to easily test these SNP frequencies as real, and new red clover material to perform this test on, is needed. If not interested in any other places in the genome then CAPS is a useful tool to genotype individual red clovers. CAPS, short for *cleaved amplified polymorphic site*, is a method using PCR-amplification (*polymerase chain reaction*) of the area where the SNPs are found. After a great number of copies have been amplified, a restriction-enzyme is used to cut the amplified DNA, but only if it recognizes a specific version of the SNP. This knowledge is used while running the amplified DNA-samples on an agarose gel. If the amplified DNA-sequence contain the special version of the SNP, the gel will show two, cut and smaller sequences which have moved further on the gel. If the amplified sequence doesn`t contain the SNP the gel will only show a single sequence, which won`t have gone as far as the cut sequences. If a gel is showing only one uncut sequence, or only the two cut sequences, it means the individual was a homozygous for one allele, but if the gel shows both the uncut sequence and the cut sequences it means that the individual is heterozygous.

Fig 1. *The ATA-restriction enzyme only cut the sequence ATA, meaning that sequence (1a) is cut, but (1b) isn`t. The T in ATA in (1a), and the middle A in AAA in (1b) are allelic versions of the same SNP. If an individual has only the T-version of the SNP, a gel will look like the one in (2a), but if the individual only has the A-version, it will look like as in (2b). If an individual is heterozygous for the SNP, it will look like as in (2c)*



To sum it up, there are two things that will be done in this experiment:

- The use of statistical method for analysis to find SNPs whose allelic versions appear with different frequency between pure stands and mixed stands of red clover.
- Verify the difference in allele-frequency in the SNPs I found by genotyping descendants of the first plants with CAPS-markers, and the descendants of some individuals from pure stands and mixed stands from another location.

2. Material and Methods

2.1 Extraction and Genotyping of DNA

In October 2012 leaf samples were collected from individuals that had survived since being sown in 2010, and since they are survivors they must have been through selection. These red clovers are further described in the experiment [10]. These samples were collected as pools of leaf samples from single individuals, 1 leaf pr. individual, 100 in total for each "pool-sample". In total, we have genetic material from 8 different plots, 4 pure stands and 4 mixed stands. Pool-sample 1-8 were from survivors in pure stands with only red clover, while 9-12 were from mixed stands. Sample 3-5 are replicate samples taken from plot 73 (number of the plot the tissue-samples were taken from), and 6-8 are replicate samples from plot 146. These tissue-samples were kept frozen.

When extracting DNA from the tissue-samples, the leaf-samples were first made into a fine powder by using liquid nitrogen together with a mortar and a pestle. DNA was then extracted as described in DNeasy® Plant Handbook [11]. After the DNA was extracted it was kept frozen until it was sent to BGI in China for genotyping-by-sequencing to identify SNPs and to obtain read-data on each allelic version of the SNPs. BGI first put the data they found through some filtration processes:

1. *Removing all SNPs from the dataset that is absent in more than half the samples, where absent means that there are less than 10 reads in the sample*
2. *Remove all SNPs with a minor-allele-frequency less than 0.05, where all samples were treated together to calculate the MAF*
3. *Removing the SNPs where the total number of reads over all 12 samples are outside the interval 100-10 000*

The data received from BGI contained 129 661 SNPs. These were either located on a known chromosomal location, 44 421 of them, with both chromosome and locus known, or on a scaffold, which was the remaining 85 240 SNPs.

2.2 Filtration of the BGI dataset, and PCA

Before any real analysis is performed on the dataset from BGI it needs to go through a couple of filtration-processes, since not all the data from BGI necessarily are valid or "good data". The first filtration used on the dataset is what in this text is referred to as a MAF0.05 filtration. MAF stands for minor-allele-frequency, and the filtration-process removes all SNPs with a minor allele frequency less than 0.05 in one or more of the samples. Our dataset consists of 12 pool-samples, and each of these have their own frequency for the ALT-allele and the REF-allele. The REF-allele is defined as the version of the SNP that is found in the databank when our genotyped sequence is compared to the sequence in the databank. The ALT-allele is the other allele. Then the filtration was performed. If the minor-allele-frequency was below 0.05 for any SNP in one or more of the 12 samples then the SNP was removed from the dataset. After this filtration was done, then all the SNPs was removed from the dataset that didn`t have a number of reads within the interval 100-499 in one or more of the samples. The sum of the number of ALT-alleles and REF-alleles for a plot within a SNP is referred to as the total number of reads it has. With this filtration method all SNPs from the dataset was removed if they didn`t have 100 or more reads or 499 or less reads in one or more of the 12 samples.

After filtration of the dataset was done, a principal component analysis (PCA) was performed, both so I could study how much allele frequency varied between pure

stands and mixed stands, and between replicates from the same plot. This was done with the software *Unscrambler* [12].

2.3 Statistics to find SNPs that appear with significant different frequencies between pure-stands and mixed-stands

F_{ST} is what I calculated and used to identify SNPs whose allelic versions appear with different frequency in pure stand -and mixed stand red clover. F_{ST} is a number calculated based on allele-frequencies and difference in allele-frequencies and is a measure of the genetic differentiation over subpopulations [6]. F_{ST} will be calculated by comparing each single plot within the pure-stand against the average of the mixed-stand, and vice versa. As mentioned in **2.1**, we got 4 pool-samples collected from four different pure stands, and 4 pool-samples collected from four different mixed stands. This will result in a total of 8 different F_{ST}-values. F_{ST} is then calculated by comparing a single stands ALT-frequency (or q-frequency) with the average ALT frequency of the other kind of stands, for example pure stand sample one is compared with the average of the mixed stands.

Equation 1. Equation for calculating F_{ST}

$$F_{ST} = \frac{\overline{q^2} - \overline{q}^2}{\overline{q}(1-\overline{q})}$$

However, F_{ST} doesn`t say anything about if the differences in allele-frequency are significant. A Chi-square test [6] are what was used to determine which SNPs that show significant difference in frequency between pure stands and mixed stands. When Fst is calculated it is used to calculate the corresponding Chi-squares. A chi-square test was then used to identify significant F_{ST}`s at different P-levels, 0.1, 0.05 and 0.01. To do this we used the test statistic $x^2 = 2NF_{ST}$, where 2N=the sum of genotyped gametes in the two populations [6]. N=100. Only the SNPs where all of the 8 F_{ST}-values were significant were themselves considered to show significant

difference in allele-frequency between pure stands and mixed stands. After the CHI^2-test FDR (*false discovery rate*) was calculated as shown in **Equation 2.**

Equation 2. Equation for calculating false discovery rate (FDR) for each P-level of the CHI^2-test. X is the number of SNPs tested in the CHI-test, P is the P-value of the test, while Y is the number of SNPs discovered to be significant.

$$FDR = \frac{X * P^4}{Y}$$

FDR is the rate at which a test will consider something to be significant when it in reality isn`t, just seem significant because of random effects. We are only interested in SNPs with known chromosomal location, so when FDR was calculated this was done with only the chromosomal SNPs, for both X and Y in **Equation 2**, and none of those located on scaffolds. I use P^4 instead of only P because I test four independent samples. I chose to work further with the P-level = 0.1 results and made chromosome map using a software called Mapchart [13], showing where and on which chromosomes the SNPs are located.

2.4 Correlation analysis

An estimate for how much allele-frequency varies between samples from the same plot is necessary when an estimate for the validity for results is needed later. After the data-filtration was done earlier, the data for each repetition of both plot 73 -and 146 was used to calculate an estimate for the variation in the allele-frequency estimated from replicate samples of a plot. Both plot 73 -and 146 each have three repetitions. The correlation coefficient (R^2) was calculated for all possible pairs of replicates (1&2, 1&3 and 2&3), and then the average R^2 for each plot.

2.5 Attempt to verify allele frequency differences in other plant material

It`s relevant to be able to verify the allele-frequency differences. This will be done by genotyping the descendants of the original survivors described in [3] and [10]. The offspring populations from survivors from a similar experiment with pure stands and mixtures at Kvithamar [10] will also be genotyped. Both the Ås and Kvithamar - descendants had been made earlier by crossing individuals within each plot with each other using bumble-bees. How many from which population, and which population was pure stand or mixed stand, and which population descended from Ås or Kvithamar is shown in **Table 1**. The descendants were grown individually and could be identified by the population they belonged to/descend from. In this project the genotyping was done by developing CAPS-markers (cleaved amplified polymorphic site). CAPS-markers can identify an individual's genotype by only cutting one of the SNPs variants.

Since the sequence with the SNP needs to be amplified to be visible on a gel, primers are developed for each SNP so a PCR for the relevant part of the sequences can be run. The primers are found with the use of a webtool called *Primer3web* [14,15,16]. Restriction-enzymes is needed to recognize and cut the amplified sequences. These were found with a webtool called *NEBcutter V2.0* [17,18]. Both primers with enzymes found can be seen in at **Table 14** in the appendix.

Table 1. *The greenhouse-populations that I collected leaf-samples from, and how many individuals that were collected from each of the populations. It was collected from 5 individuals from each of the populations 1-5, 6 from population 6, 17 from population 73 and 16 from each of the populations 84, 146 and 150.*

Populations	Pure or mixed	From Location	Number of individuals per population
1-3	Pure	Kvithamar	5
4-5	Mixed	Kvithamar	5
6	Mixed	Kvithamar	6
73	Pure	Ås	17
146	Pure	Ås	16
84, 150	Mixed	Ås	16

The extraction of DNA from the greenhouse-individuals were done with an *DNeasy 96 Plant kit* [11] *from Qiagen*, and since the samples had been stored at -80^{O}C, the instruction "Protocol: Purification of Total DNA from Frozen or Lyophilized Plant Tissue" [11]. After the extraction of DNA was done, the quality of the DNA was tested on a 1% agarose gel-electrophoresis. To check the concentration in our samples they were studied by using a spectrophotometer. The samples were then stored in a freezer awaiting to be genotyped.

Before any actual genotyping can start, a couple of checks will be performed with some old DNA-samples, the very same DNA samples sent to BGI at the start of this project. The goal now is to perform a test on the different primer-pairs acquired to see if their PCR-processes make enough DNA to create a clear gel-image. A primer-pair is discarded if the gel-electrophoresis-image is weak, or if it is a little "smeared", or shows several bands, which will mean that the primers or the PCR-program doesn`t work properly. A master PCR-mix is made to PCR-amplify the sequences with the SNPs within them.

Table 2. *The recipe for the PCR reactions.*

Reagent	**Pr. Reaction (µL)**	**Concentration**
Jumpstart™ -mix	10	1x
Primer L (10µM)	0.8	0.4 µM
Primer R (10µM)	0.8	0.4 µM
Milli-Q water	7.4	
DNA-template	1	
Total	20	

The PCR-mastermix itself only contain the three first ingredients, Jumpstart™ -mix, [19] both the primers and milliQ-water. The mix is made by multiplying the "pr.

reaction" volumes of Jumpstart-mix, both primers and milliQ-water as mentioned in **Table 2** according to the equation below:

#DNA-samples + #positive controls + #negative controls + (1 or 2)

The +(1 or 2) is just to make sure that enough PCR-mastermix is made. This mix in short contain rich doses of the different nucleotide bases, taq-polymerases and the relevant primer-pairs. 19 µL of this master-PCR-Mix is added to a set of tubes. These tubes then have 1 µL of either BGI-sample 3, 4, 5, 8, 9 or 12 added to them, for a total reaction volume of 20 µL. Negative controls with milliQ-water instead of DNA was also made.
The samples were then put in a pre-programmed PCR-machine.

Fig 2. *The programmed steps for the PCR-machine. Step 2-4 are repeated 35 times.*
Step 1: 94^{o}C for two minutes –(denaturation)
Step 2: 94^{o}C for 30 seconds –(denaturation)
Step 3: 55^{o}C for 30 seconds –(annealing)
Step 4: 72^{o}C for 30 seconds –(extension)
Step 5: 72^{o}C for 5 minutes –(end extension)
Step 6: 4^{o}C ∞ -(waiting)

Step 2-4 in **Fig 2** are repeated 35 times. For each cycle in the PCR-machine, the quantity of DNA in the samples increase exponentially. The annealing temperature might need to be adjusted according to which pair of primers that are being used, since the optimal annealing temperature varies with primer pairs.

Eight different primer-pairs were tested, and not all of these had an optimal annealing temperature at 55^{o}C. Some of the primer-pairs therefore went through PCR-multiple times but with different annealingtemperature. The final annealingtemperature can be seen in appendix **Table 17**. These final annealingtemperatures were used later when doing the PCR for the genotyping of the Vollebekk individuals. A 2% agarose gel-electrophoresis was then run with the

PCR-products. After all the primer-pairs were tested, only a few of them went further (**Table 18** in the appendix) to what is referred to as “test-cutting” or “test-digestion” with restriction-enzymes. These primer-pairs gave a clear gel-image after PCR and is now going to be tested for if the genotypes are easily distinguishable after the PCR-products have been digested by their relevant restriction-enzyme. An enzyme-mix was made.

Table 3. *The cut-mastermix. Both the enzyme during the mixing, and the cut-mastermix should be kept cool on ice while not in a freezer.*

Reagent	**Pr. Reaction (µL)**
Buffer	2
Enzyme	0.2
Milli-Q water	7.8
PCR-product	10
Total	**20**

The cut-mastermix was made by multiplying the “pr. reaction”-volume of buffer, enzyme and milliQ-water with the same number as for the PCR, but without the negative controls.
10 µL of this was added to 10 µL PCR-product. The restriction-enzyme digestion reaction was performed at temperatures recommended by the manufacturer, and for 15 minutes. More details in Appendix **Table 18**. After the enzyme digestion was done the samples were run on a 2% agarose gel-electrophoresis.
Which primer-pairs with restriction-enzyme that passed the PCR -and cutting test is shown in **Table 19** in the appendix.

After the testing for primer-pair with enzyme was done the actual genotyping process could begin. Both PCR and restriction-enzyme digestion are now performed on the individuals from the Vollebekk Greenhouse, but pretty much the same way it was performed earlier. However, there is some differences. These are described in the **appendix 5.5**:

- Run on gel only after digestion with restriction-enzyme, not after PCR.
- PCR-product for TP2_18520944 F4R4 for column 5-12 were digested for 30 min.

After being put through a 2% agarose gel-electrophoresis, the individual samples on the gel-images was interpreted as either one of the two homozygotes, or the heterozygote.

3 Results

3.1 BGI-data

After BGI had sequenced our DNA samples, a total of 129661 SNPs was found. Each SNP was given with a chromosomal location (chromosome+locus) if they were located on a chromosome and not a scaffold, the number of reads containing our «reference nucleotide» (REF) and the number of reads containing the alternative nucleotide (ALT). This was given for each of the 12 samples.
On the data I received from BGI I filtrated away all the SNPs that didn`t have a sum of reads within the interval 100-499 in one or more of the samples. I also removed all the SNPs that had a minor-allele-frequency less than 0.05 in one or more of the samples. Both the filtration-processes were done with each replicate within 73 and 146 counting as a single sample.

3.2 PCA

Minor allele frequencies (MAF) was calculated for our plots 1-12, and then put through a principal component analysis (PCA) in a software called *Unscrambler* [12]. The result is shown in **Fig 3.**

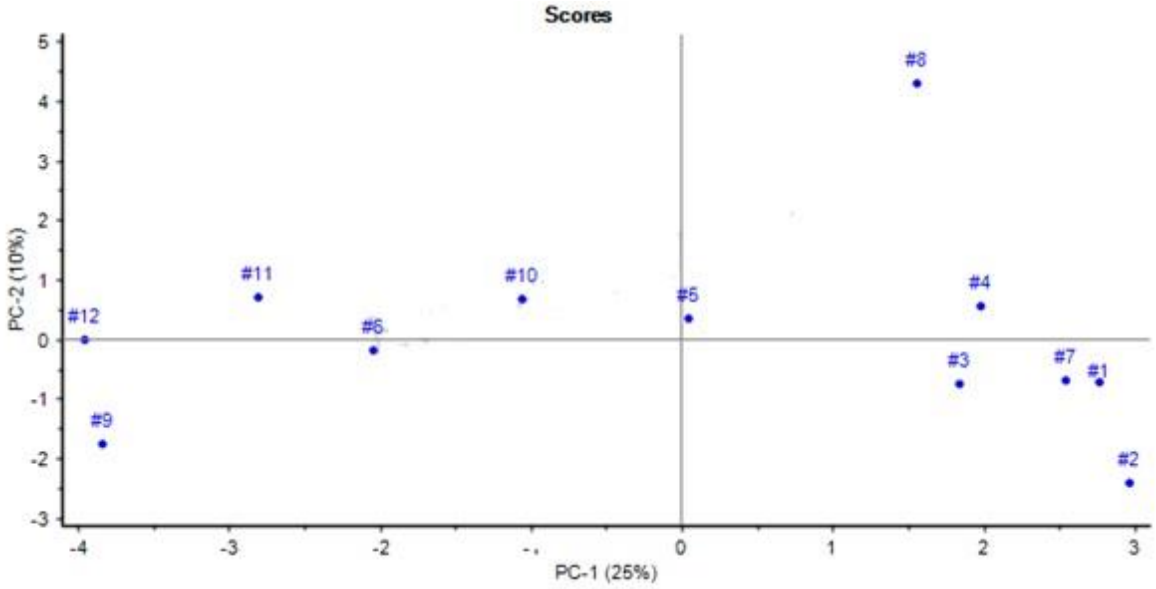


Fig 3. *Principal component analysis after removing SNPs with a number of reads outside of the interval 100-499 in or more of the samples and removing all SNPs with a MAF less than 0.05 in one or more of its samples. The PCA is based on the calculated minor allele frequency for all plots 1-12. Plot 1-8 are the pure-plots, and 9-12 are the mixed-plots. Plot 3-5 are replicates of the same plot (73) and 6-8 are replicates of the same plot (146).*

The first axis of the PCA, the PC-1 axis, explains 25% of the variation among the plots. As mentioned earlier, stand 1-8 was the pure stands, and 9-12 was the mixed stands, and if we study how the stands are distributed along the PC1, it seems that all the mixed stands are on the left side of PC1, and almost all pure stands, except 6, is on the right side. Short, PC1 seem to divide the plots into groups of pure stands and groups of mixed stands. PC2 explains about 10% of the variation among the plots but doesn`t have a visible pattern as PC1 have. PCA is a kind of a statistical analysis method were samples are compared to each other and shows of much they vary compared to each other. The closer some samples are to each other, the less they vary compared to each other. Since the axes explains different amounts of variation, this should be considered when studying distance between samples.

3.3 Correlation between replicates

In **Fig 3** above, plot 3-5 are the 73-replicates, and 6-8 are the 146-replicates. It is hard to see any correlation between the 146-replicates, because these are far apart. The 73-replicates are closer to each other, which might indicate a stronger correlation between the 73-replicates.

3.4 Comparison between pure plots, and mixed plots

The pure samples are sample 1-8 in the PCA, while the mixed samples are the samples 9-12. In the PCA these two groups seem to be divided into two groups, one with only pure samples, and one with almost only mixed samples. The pure samples seem to mostly cluster together, with sample 5 somewhat far away. Sample 6 is in the group of “mixed samples», indicating it to be different from the other pure samples. However, this is caused by random variation, since individuals from the same plot have been under the same kind of selection pressure relative to the SNPs we are researching and should therefore not have any other variation than what is caused by random effects. As mentioned, the mixed samples seem to be grouped to themselves but compared to the pure samples group, the mixed samples doesn`t seem to cluster as much, and is more spread, indicating that there is more variance within the mixed group than within the pure group.

3.5 Comparison of the replicates

We have already done a PCA to analyze the correlation between replicates. Another method is to calculate the average of the R^2 for each possible relationship (1-2, 1-3, 2-3), and the average R^2.

Table 4. *Comparative statistics for the replicates from plot 73 and plot 146, based on data extracted from the whole dataset, which had been MAF0.05 filtered and read100-499 filtered.*

	R^2 1_2	R^2 1_3	R^2 2_3	The average R^2
73	0,90	0,89	0,89	0,90
146	0,87	0,87	0,88	0,87

3.6 Identification of SNPs with significantly different allele frequencies in pure stands vs. species mixtures

After removing the SNPs that had one or more samples with a number of reads outside the interval 100-499 and removing all SNPs with one or more minor-allele-frequencies below 0.05 it`s necessary to know which of them that appear with a significant different allele-frequency between the pure stands and the mixed stands. To do this we used a CHI^2-test. To use this test, I first calculated the F_{ST}-values. We have twelve samples, but since 3-5 are replicates of stand 73, and 6-8 of stand 146, I have in reality 4 pure stands and 4 mixed stands. This means that an average frequency was calculated for the 73 replicates, and the 146 replicates. The F_{ST}-values were calculated by comparing the q-frequency (ALT-frequency) of every single pure stand against the average of the mixed stands and vice versa. This way each SNP have eight different F_{ST}-values, and each of these for all the SNPs was tested with a CHI^2-test at p-level 0.1, 0.05 and 0.01. All the SNPs that didn`t show significance in all its F_{ST}-values was removed from the dataset, giving us three datasets, one for each p-level. Some of the SNPs were located on a known chromosomal locus, but others were only located on scaffolds. The ones located on scaffolds (85 240 SNPs) were removed from the datasets, since many of them appeared to be in highly variable regions. At P-level = 0.1 it was nine different SNPs that showed a significant different allelic variation between pure stands and mixed stands. The number was reduced to six for p-level = 0.05, but none were left at p-level 0.01. False-discovery-rate was also calculated for each p-level. Which SNPs that showed significant allelic variation between pure stands and mixed stands at different p-levels, and their FDR, are shown in **Table 5** below.

Table 5. *Chromosomal SNPs calculated to be significant at the p-levels 0.1, 0.05 and 0.01 when our data have been put through a MAF0.05 and read100-499 filtering. An "X" marks which of the SNPs are significant at the different p-levels. There is no SNPs with a known chromosomal location that is significant at p-level 0.01. At the bottom of each column is the calculated FDR for the P-level*

	SNP significant when p-level =0.1	SNP significant when p-level = 0.05	SNP significant when p-level = 0.01
Tp2_18520944	x	x	
Tp2_29283780	x	x	
Tp4_11383593	x	x	
Tp4_11383594	x	x	
Tp4_23347267	x	x	
Tp4_23347286	x	x	
Tp4_24549667	x		
Tp5_10561679	x		
Tp6_4397690	x		
FDR	0,05	0,005	-

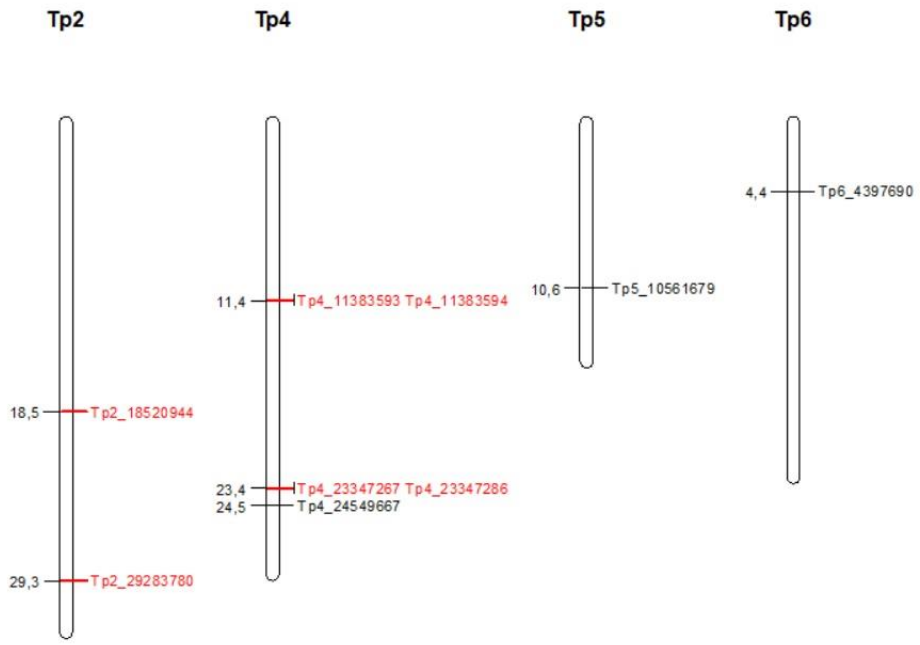


Fig 4. *Chromosomal maps for the SNPs we discovered to appear with significant different allele-frequency between the pure-stands and the mixed-stands. The ones that are red are significant at both p-level = 0.1 and 0.05, while the black ones are only significant at 0.1. The map was made with a software called MapChart* [13].

Table 6. The average q-allele (ALT) frequencies and SEs in both pure stands and mixed stands for all the nine SNPs tested to be significant in a CHI^2-test at P-level = 0.1.

SNP	Average q-allele freq. in the pure stands	Average q-allele freq. in the mixed stands	SE for q-allele freq. in pure stands	SE for q-allele freq. In pure stands
Tp2_18520944	0,15	0,34	0,01	0,03
Tp2_29283780	0,55	0,24	0,04	0,04
Tp4_11383593	0,46	0,20	0,09	0,04
Tp4_11383594	0,46	0,20	0,09	0,04
Tp4_23347267	0,44	0,70	0,03	0,04
Tp4_23347286	0,44	0,70	0,03	0,04
Tp4_24549667	0,53	0,71	0,01	0,04
Tp5_10561679	0,25	0,11	0,03	0,01
Tp6_4397690	0,28	0,46	0,02	0,02

3.7 Development of CAPS-markers

After finding out which SNPs that seem appear with significant allelic variation between pure -and mixed stands an effort were put into finding out if they appear with about the same frequency in another set with red clovers. To do this CAPS-markers were developed and used, and the new material that were genotyped was descendants of the survivor-generation sent to BGI for GBS. First, I needed to find primers that could be used to PCR-amplify the areas where the SNPs were located. Possible primers were found for four SNPs deemed to be significant at p-level= 0.1. Only a few of them however were chosen for testing. I chose the four with the greatest difference in allele frequency between pure stands and mixed stands.

- TP6_4397690
- TP4_23347267
- TP4_11383593
- TP2_18520944

These were tested by using them to PCR-amplify, and then having their gel-image studied after the gel-electrophoresis was done. These results can be seen in the appendix, **Gel1 – Gel23**.

With the restriction-cutting done, each sample on the gel-images need to have their genotypes interpreted. Each individual will have one of three genotypes, aa, bb or ab. In this instance aa means homozygote uncut, bb is homozygous cut, and ab means heterozygous cut and uncut. Cut and uncut is different allelic versions of the SNP. How these different genotypes look like on a gel is shown in **Fig 5** beneath.

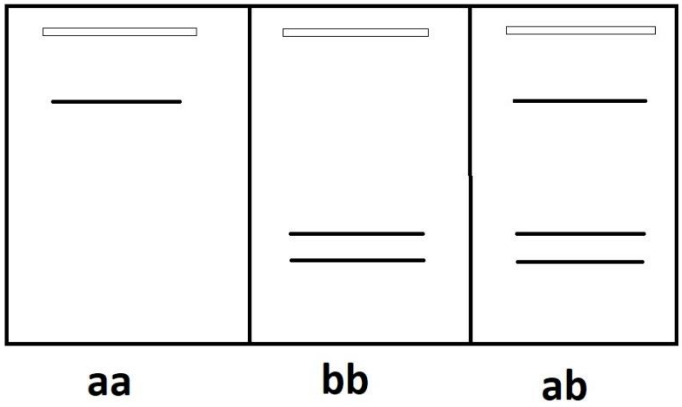


Fig 5. *To the far left is the gel-image for the aa "homozygous uncut" -genotype, while the middle shows the bb "homozygous cut" -genotype. To the far right of the image shows both an uncut sequence and cut sequences, meaning this is the ab "heterozygous cut and uncut" -genotype. This image was used when interpreting the gel samples*.

The result of the interpretation of the gel-images for TP2_18520944 and TP6_4397690 can be seen in **Table 7** and **Table 8** respectively.

Table 7. *How many individuals within each population that had the different genotypes*. *The SNP is TP2_18520944. The "blank" genotype is for the individuals that didn`t show anything on the gel-image. The a-allele = the ALT-allele, b-allele = the REF-allele.*

	1	2	3	4	5	6	73	84	146	150
aa	0	0	0	0	0	0	0	0	1	0
b*b*	3	3	3	4	4	5	16	11	8	11
ab	2	2	1	0	1	1	1	4	7	5
Blank	0	0	1	1	0	0	0	1	0	0

Table 8. *How many individuals within each population that had the different genotypes*. *The SNP is TP6_4397690. The "blank" genotype is for the individuals that didn`t show anything on the gel-image. The a-allele = the ALT-allele, b-allele = the REF-allele.*

	1	2	3	4	5	6	73	84	146	150
aa	3	1	0	2	1	1	4	8	7	3
bb	0	2	1	1	1	1	5	5	2	5
ab	2	2	3	1	3	4	8	2	6	8
Blank	0	0	1	1	0	0	0	1	1	0

The data presented in **Table 7** and **Table 8** can be used in further analysis. For example, we know that individuals descended from population 1-6 are from Kvithamar and the rest is from Ås. We also know that population 1-3, 73 and 146 are pure-stands red clover, while population 4-6, 84 and 150 are mixed-stands red clover. Various kinds of allele frequencies can be found. The general a-frequency and b-frequency, a -and b frequencies with pure -and mixed stands, a -and b frequencies within the groups Ås-individuals and Kvithamar-individuals, and frequencies within combination-groups such as Ås-pure stands or Kvithamar-mixed stands.

Table 9. *Allele frequencies, for TP2_18520944 F4R4, within each group of pure or mixed, and Ås or Kvithamar. The a-allele = the ALT-allele, b-allele = the REF-allele.*

	a-freq.	b-freq.
Pure	0,16	0,84
Mixed	0,12	0,88
Ås	0,14	0,86
Kvithamar	0,12	0,88

Table 10. *Allele frequencies, for TP2_18520944 F4R4, within each combination of location and type of growth culture. The a-allele = the ALT-allele, b-allele = the REF-allele.*

a-freq	Pure	Mixed
Ås	0,17	0,11
Kvithamar	0,18	0,07
b-freq.	Pure	Mixed
Ås	0,83	0,89
Kvithamar	0,82	0,93

Table 11. *Allele frequencies, for TP6_4397690 F3R3, within each group of pure or mixed, and Ås or Kvithamar. The a-allele = the ALT-allele, b-allele = the REF-allele.*

	a-freq	b-freq
Pure	0,53	0,47
Mixed	0,52	0,48
Ås	0,53	0,47
Kvithamar	0,52	0,48

Table 12. *Allele frequencies, for TP6_4397690 F3R3, within each combination of location and type of growth culture. The a-allele = the ALT-allele, b-allele = the REF-allele.*

a-freq.	Pure	Mixed
Ås	0,53	0,53
Kvithamar	0,54	0,50
b-freq.	Pure	Mixed
Ås	0,47	0,47
Kvithamar	0,46	0,5

Table 13. *The table shows what the a-allele frequency was in the GBS data, and what it was in the Vollebekk greenhouse genotyping.*

a-allele frequencies	**Pure stands GBS-data**	**Pure stands CAPS-assay**	**Mixed stands GBS-data**	**Mixed stands CAPS-assay**
TP6_4397690	0.28	0.53	0.46	0.52
TP2_18520944	0.15	0.16	0.34	0.12

4 Discussion

4.1 Pooled DNA-samples

The DNA-samples sent to BGI for genotyping-by-sequencing (GBS) were pooled, meaning instead of sending a lot of single individuals DNA-samples, we instead choose to send a total of 12 different samples, which each was a pool of DNA-samples from 100 different individuals. Ultimately, a genotyping of all the single individuals would give more secure data, but according to [21], pooling the DNA from many individuals this way, followed up by a filtration of the data received from the GBS can reach an accuracy that is above 90% of what we get from genotyping every single individual. Considered time and resources saved by genotyping pooled DNA-

samples instead of single individuals, a drop of only a few percent accuracy is acceptable.

4.2 Filtration of the dataset

The dataset received from BGI contained several things. It had all the SNPs in *T. pratense* they had managed to identify by GBS, which chromosome or scaffold they were located on, and their locus if they were on chromosomes. For each SNP they had identified what is referred to as a REF-allele (reference-allele) and an ALT-allele (alternative allele), and how many "reads" was detected for each of these. All 12 pooled samples had their own number of ALT-alleles and REF-alleles. The REF-alleles is the version of the SNPs they found that was the same as what is shown in the LIS database [20], while the ALT-alleles are not. As mentioned above, this dataset need to be filtered if we are going to be able to use it. One way to filtrate the dataset to increase accuracy is to remove all SNP with a total number of reads below 100 in one or more of the 12 samples. The reason for this is simple, we need to reduce the impact of single observations. With just a few reads the estimated allele-frequency becomes more uncertain. Another related filtration performed was the read499, meaning that all SNPs with a total number of reads at 500 or above in one or more of the 12 samples is removed. This seems strange at first, as more data usually means more accurate statistics. However, according to [21], it seems that accuracy in data decrease when reaching 600 reads and above. Here they had compared data gained from genotyping individuals with data gained from genotyping pools with individuals, and the statistical accuracy for SNPs with 600 and above number of reads seems to drop from around 90% to around 80%. To keep a certain level of accuracy it was therefore decided to keep the number of reads in the interval 100-499.

Another kind of filtration our SNPs went through is what we refer to as the MAF0.05-filtration. Details for how the MAF is defined can be found in **method and material 2.2**, but with the MAF0.05-filtration method all SNPs with a MAF below 0.05 in one or more of the 12 samples are removed. This is because low allele-frequencies are very uncertain. For example, if we by random sampling of 100 individuals get a frequency of 0.01, we can`t be certain if this in reality is 0.001, or maybe 0.02. Because of this uncertainty we remove the SNPs with one or more MAFs at 0.05.

4.3 PCA

After the filtration of our dataset was done I performed a principal component analysis on the MAF-values. The results can be seen in **Fig 3**. It`s easy to think that plots in a PCA that is far apart is very different, but remember to consider the scale of the axis, and from which axis the viewpoint is. In our PCA sample 3 and sample 8 might seem far away from each other but is actually very close if only viewed from the X-axis. When studying the different plots according to the PC1, it seems to distinguish the plots into two "groups", a group with pure stands (samples 1-8) and a group with mixed stands (samples 9-12). However, the pure -and the mixed stands are not perfectly divided, because stand 6, which is a pure stand, is grouped with the mixed stands on the PC1 (X-axis). Why stand 6 is among the mixed stands might be because of random effects in the 100 individuals making up stand 6, combined with the fact that PC1 only explain about 25% of the total variation in our 12 pooled samples.

PC2 explains about 10% of the total variation between the 12 samples, but compared to PC1, it`s hard to define any groups based on any data we have about our samples. The experiment is design to only have two distinct groups with *T. pratense*, pure stands and mixed stands, and beyond PC1 they don`t seem to be any different.

4.4 Analysis of the replicates

To know something about the variation between distinct groups and plots we also need to study the variation within a plot. If the variation within a plot is great, then the variation between plots might not be as significant as it seems. To study the within-plot-variation we took three replicate samples from each of two plots. We took three samples with 100 individuals each from plot 73, and the same for plot 146. Sample 3-5 in the 12 samples are the plot 73 replicates, while sample 6-8 are the 146 replicates. All these samples were put through PCA with the rest of the samples. When studying the PCA-plots for the 73 replicates we can see that there is very little variation between 73-replicate 1 and 2 (sample 3 and 4), while replicate 3 (sample 5) are some further away, indicating some variance. In total the 73-replicates seem to vary very little, indicating only a low level of variation in stand 73. However, when

studying the PCA of the 146-replicates it seem to be a great variation, as each replicate is far away from one another. Replicate 2 and 3 (sample 7 and 8) are not that far away from each other on PC1, but there is some distance between them on PC2, indicating some, but little variance. Replicate 1 (sample 6) are however far away from both replicate 2 and 3, showing a great variation within plot 146.

PCA is somewhat only a visual representation of the variation between replicates of the same plot, we can also calculate R^2 as an estimation of the correlation between the replicates of the same plot. The result of this calculation is shown in **Table 4**. The 73-replicates have an estimated R^2 around 0.9, meaning that there is a very high correlation between the 73-replicates. The 146-replicates have R^2s just slightly below 0.90, which also is very good, only a few per cents away from what it was in the 73-replicates. There is some variation, but just little. The variation within each group of replicates (73 and 146) is due to random effects, since they have been subject to approximately the same selection pressures. The greatest variation in the MAF-dataset seem to be between pure stands and mixed stands. However, there is also a very large variation between replicate samples from the same plot. As mentioned, this is due to random effects. It seems to be a large variation between the mixed stands, but less among the pure stands. All in all it`s a larger variation between pure stands and mixed stands than there is among replicate samples of the same plot, but not much.

4.5 Discovery of significant SNPs

To find which of the SNPs in the list that had significantly different allele frequency in the pure stands and the mixed stands, I first calculated a pairwise F_{st} between the single pure plots and the average of the mixed stands and vice versa, and then put them through CHI^2-tests with the P-values 0.1, 0.05 and 0.01. The result of the CHI^2-test can be seen in **Table 5**. When performing the CHI^2-test I only consider the SNPs with a known chromosomal location, and not those located on scaffolds. At the least stringent test level, P = 0.1, I get a total of 9 different SNPs that are considered to appear with significant different frequency between the pure stands and the mixed stands. When the stringency level is increased to P = 0.05, we are left with 6 different SNPs, but when I increase the stringency further to P = 0.01 we got no SNPs. For

obvious reasons we can`t work further with P = 0.01, so we must choose between P = 0.1 and P = 0.05. Considered how few SNPs we are left with we find it best to work with P = 0.1, which also gives us an acceptable FDR = 0,05 (**Table 5**), meaning that about 5% of our SNPs seem significant due to random effects, without actual being significant. What the actual allele-frequency for these SNPs are is shown in **Table 6**. Here we can see a clear difference between the average frequency of the pure stands and the mixed stands, with some SNPs having a greater frequency within the pure stands, and others having a greater frequency within the mixed stands. These frequencies become relevant later when we develop a marker-assisted method to read an individual's genotype for specific SNPs. The standard error for each stands SNP-frequency is also at an acceptable level. This mean that if we take the average q-allele frequency (ALT-frequency) ± 2*SE (i.e. the 95% confidence interval) for both the pure stands and the mixed stands the intervals won`t overlap.

4.6 Genotyping

The individuals tested to be significant at P = 0.10 is the basis for our work further to develop CAPS-markers to genotype DNA-samples collected at Vollebekk greenhouse. Out of these only four SNPs were tested, and only two of the SNPs had a CAPS-marker fully developed for them, TP6_4397690 and TP2_18520944 (name = chromosome + locus). Most of the individuals that we genotyped were descendants of survivors from the population we original sent to BGI for GBS. In addition, the offspring populations from survivors from a similar experiment with pure stands and mixtures at Kvithamar, were tested. The descendants are hypothetically not too different in allele-frequency than their parent-generation, but some differences are to be expected. After genotyping the greenhouse individuals by interpreting our 2%-agarose-gels, we calculated the allele-frequencies for both pure-stands and mixed-stands, and for whether they were from Ås or Kvithamar (the parents were collected from both Ås and Kvithamar). I was hypothesizing that the frequency for at least the pure stands and mixed stands would be different from each other, but as can be seen in **Table 9** and **Table 11** there isn`t much difference in allele-frequency in neither TP6_4397690 or TP2_18520944. Not only were the difference very small, it was almost non-existent. For TP6_4397690 the difference in the "a-allele" was only 4%,

and for TP2_18520944 it was only 1.1%. It`s only in the parent/survivor-generation there is a significant difference in the frequency of the allelic versions of the SNP. This might be because bumble-bees were used to pollinate the parent-generation, and the bumble-bees might have been selective, and not as random as we hypothesized, causing a distortion from the allele-frequencies I hypothesized. Another possible source of influence on the allele-frequencies is the variation in the number of seeds each maternal plant in the parent/survivor-generation produced, causing an uneven contribution to the next generation, breaking one of the Hardy-Weinberg laws for maintaining the allele-frequencies from one generation to the next. It`s also possible that my CAPS-assays didn`t work optimally. There were some problems with uncomplete digestion, making it hard to differentiate between genotypes on the gel. Some samples even had “double fragments”, showing two fragments even though it wasn`t one of the ordinary cut sequences.

5. Conclusion

With filtration and testing of the dataset received from BGI I found 9 SNPs whose allelic versions seem to appear with different frequency in pure stand red clover and mixed stand red clover, this with an FDR = 0.05. However, when I tried to get this confirmed by genotyping the descendants of the original plants genotyped, and the descendants of the survivors from a similar experiment on Kvithamar, I was met with very different frequencies for the SNPs allelic versions. There was no significant different frequencies between the SNPs in the pure stand descendants and the mixed stand descendants. My conclusion is that there are allelic versions of some SNPs that appear with different frequency in the pure stand red clover and the mixed stand populations sequenced by GBS, but I was not unable to confirm this in the other populations tested.

6. References

1. Long, S.R., Staskawics, B.J., (1993) Prokaryotic Plant parasite. *Cell volume 73, Issue 5*. Pages. 921-935.
2. Younie, D. (2012) Grassland management for organic farmers. The Crowwood Press Ltd. ISBN 978 1 84797 387 0
3. Ergon, Å., Kirwan, L., Bleken, M.A., Skjelvåg, A.O., Collins, R.P., Rognli, O.A. (2016). Species interactions in a grassland mixture under low nitrogen fertilization and two cutting frequencies: 1. dry-matter yield and dynamics of species composition. *Grass and Forage Science 71, 667-682.* doi: 10.1111/gfs.12250
4. Peyraud, J.L., Le Gall, A., Lüscher, A., (2009) Potential food production from forage legume-based-systems in Europe: an overview. *Irish Journal of Agriculture and food Research 48*. Page. 115-135
5. Carlsson, G., Huss-Danell, K. (2003) Nitrogen fixation in perennial legumes in the field. *Plant and soil 253*. Page. 353-372
6. Hedrick, P.W. (2011) *Genetics of Populations* (4th ed.). Jones and Bartlett Publishers ISBN-13: 978-0-7637-5737-3. ISBN-10: 0-7637-5737-3
7. Ching, A., Caldwell, K.S., Jung, M., Dolan, M., Smith, O.S., Tingey, S., Morgante, M., Rafalski, A., (2002) SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. *BMC Genetics*.
8. Milvang. Ø., (2016). Identification of regions that control timing of elongation in red clover using Genotyping by Sequencing (GBS). NMBU, Ås, Norway.
9. Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S., Mitcheel, S.E., (2011) A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLoS ONE*. doi:10.1371/journal.pone.0019379
10. Ergon, Å., Bakken, A.K., (2016) Red clover traits under selection in mixtures with grasses versus pure stands. *Grassland Science in Europe, Vol. 21 – The multiple roles of grassland in the European Bioeconomy.* Pages. 811-813
11. DNeasy® Plant Handbook (June 2015). QIAGEN®.
12. The Unscrambler® X Version 10.5. *CAMO*. (http://www.camo.com/rt/Products/Unscrambler/unscrambler.html) – 05.apr.2018

13. Voorrips, R.E., (2002). MapChart: Software for the graphical presentation of linkage maps and QTLs. *The Journal of Heredity* 93 (1): 77-78. https://doi.org/10.1093/jhered/93.1.77
14. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M and * Rozen SG. Primer3--new capabilities and interfaces. Nucleic Acids Res. 2012 Aug 1;40(15):e115.
15. Koressaar T and Remm M. Enhancements and modifications of primer design program Primer3. Bioinformatics 23(10):1289-91
16. Link to Primer3Web version 4.1.0 - http://bioinfo.ut.ee/primer3/ (15.apr.2018)
17. Vincze, T., Posfai, J. and Roberts, R.J. NEBcutter: a program to cleave DNA with restriction enzymes. Nucleic Acids Res. 31: 3688-3691 (2003)
18. Link to NEBcutter V2.0 - http://nc2.neb.com/NEBcutter2/ (15.apr.2018)
19. Jumpstart™, (2013), Product Information, *Sigma-Aldrich*
20. https://legumeinfo.org/ (24.apr.2018)
21. Ergon Å, Skøt L, Sæther VE, Rognli OA. (2018). Characterization of selection occurring within one generation of *Trifolium pratense* L. (red clover) from sowing until the end of the second harvesting year. Manuscript.

7. Appendix

Table 14. *P-values with corresponding test-statistics used in chi-testing*

P-value	0.1	0.05	0.01
Test-statistics	2.71	3.84	6.64

7.1 Finding primers

Primers are used to make multiple copies of specific sequences through the process of PCR.

When SNPs that appear with different frequencies in pure-stands and mixed-stands were found, we used their surrounding sequences, as given to us by BGI, to expand the surrounding sequence with +500 at each side of the sequence. To do this the "blastn"-function on *Legumeinfo.org* was used. This resulted in the SNPs themselves

always being at base nr. 551 in our expanded sequence. This was done for all the SNPs mentioned in **Table 5**. A webtool called “Primer3web” [14,15,16] was used to find primers. The whole sequences (500+BGI+500) were used to find primers. The webtools default settings were used for almost all SNPs. Only one SNP-sequence didn`t use the default settings. This was the “double SNP”, the sequence for the two SNPs 11383593 and 11383594 at chromosome 4. Here the sequence was modified with a [ ] around the two SNPs, as this told the webtool that these two bases needed to be included in the sequence covered by the primers.

7.2 Finding restriction enzymes

Restriction enzymes are the tools used to study the genotype of an individual.
To find restriction-enzymes we can use we use the SNPs BGI-sequence to do searches in a webtool called “NEBcutter V2.0” [17,18]. The reason only the BGI-sequence is used here is because NEBcutter only can zoom in at about 70-80 bp at maximum. Only default settings were used.
An area from about bp 20 to bp 80 is zoomed in at. We can now see single bases. The tool shows us a collection of restriction-enzymes related to this sequence and show us which part of the sequence the restriction-enzyme recognize. A list with restriction-enzymes that recognize our SNPs are made.
This list is repeated to also get a list of the restriction-enzymes that recognize the alternative SNP.

7.3 Material for further work

Instead of working with all of the significant SNPs and their primers, we choose to work with only four of them. These were chosen because they showed the greatest and most significant difference in allele-frequency between pure-stands and mixed-stands. Chosen SNPs with associated primers are in **Table14**:

Table 14. *The primers, and therefore also the SNPs I chose to work further with, their primer sequences and their restriction enzymes.*

SNP	Primer names	Primers sequences	Restriction enzymes
Tp2_18520944	Tp2_18520944_F1	CTGTTCCGAAAGCAGCAGTT	HaeIII
	Tp2_18520944_R1	TTTGCAGGCTTGAGACATGG	
Tp2_18520944	Tp2_18520944_F2	ATCGAGTTTGTGCCATGACG	HaeIII
	Tp2_18520944_R2	GTATGATGCAAACCAGCCCC	
Tp4_11383593	Tp4_11383593_F1	CCACCAGCCGAAGATGAACT	MseI
	Tp4_11383593_R1	TGCACTTCGACATCACAAGC	
Tp4_11383593	Tp4_11383593_F2	TGCACTTCGACATCACAAGC	MseI
	Tp4_11383593_R2	GCTTGCAGCGATCATTCACA	
Tp4_23347267	Tp4_23347267_F1	CAGGCACCAAAAGTCAACCA	ScaI HF
	Tp4_23347267_R1	TGGCTCCATTGGACAAGGAT	
Tp4_23347267	Tp4_23347267_F3	ATCAAACAATGACGGCAACA	ScaI HF
	Tp4_23347267_R3	CCGAGTTCACTGACACCTCA	
Tp6_4397690	Tp6_4397690_F2	AGTGACTCAGCTGGAAACGA	BstUI
	Tp6_4397690_R2	CGACATCACGTTCCGATTGT	
Tp6_4397690	Tp6_4397690_F3	GGCAACTAAGCGGTGTTAGC	BstUI
	Tp6_4397690_R3	GGCGAAATTAAGCTCAACGA	

The primers are named after their associated SNP, plus an F for the forward primer and an R for the reverse primer. The primers also work as pairs and are named accordingly to be easily recognized as pairs, for example TP2_18520944 F1 and R1 work together as a pair.

7.4 Collection of plant-material

Plant-material to test our primers and restriction-enzymes were collected from individuals kept in a greenhouse at Vollebekk, Ås. The table shows which sample is which individual from the greenhouse. The individuals will be mentioned as which tray they are grown on, and which population they are descended from. The samples were collected as sample A-H in column 1-12.

Table 15. The identity of each red clover individual. Their identity is tray-population.

	1	2	3	4	5	6	7	8	9	10	11	12
A	11-5	11-4	1-3	24-3	2-72	7-6	8-146	31-73	3-150	30-150	25-146	14-146
B	11-1	1-146	1-1	24-4	2-3	7-1	8-73	31-84	3-84	30-84	25-84	14-84
C	11-84	1-84	1-2	24-73	2-5	7-2	8-150	31-146	3-73	29-84	24-84	14-150
D	11-146	1-150	1-73	24-6	2-4	7-3	8-3	31-150	28-84	29-73	24-73	15-146
E	11-6	7-146	24-84	24-146	2-6	7-150	20-146	22-84	28-150	29-146	24-150	24-146
F	11-2	1-4	24-1	24-150	2-2	8-84	20-84	22-73	28-73	29-150	24-146	24-150
G	11-73	1-5	24-5	2-84	7-73	8-1	20-73	22-146	28-146	25-73	6-84	20-6
H	11-150	1-6	24-2	2-150	7-4	8-5	20-150	3-146	30-73	25-150	6-73	3-73

The red clover (*T. pratense*) at the Vollebekk greenhouse are descendants of survivors grown at two places in Norway, Ås and Kvithamar, in plots of pure stands (only red clover) and mixed stands with a clover/grass combination. These survivors were dug up and planted as single individuals with a flowerpot each, and then placed within the Vollebekk greenhouse. After vernalization and flowering, the plants were crossed with other plants within the same plot using bumblebees. The offspring of these crosses are what we collected leaf-tissue samples from [10].

Extraction and fixation of DNA were done with a “DNeasy 96 plant kit” [11], which meant that we could extract DNA from 96 individuals in total:

Table 16. How many individuals that were tested from each population, and which of the population that were pure stand or mixed stand, and which were from Ås or Kvithamar.

Populations	Pure or mixed stand	From location	Number of individuals within population
1-3	Pure	Kvithamar	5
4-5	Mixed	Kvithamar	5
6	Mixed	Kvithamar	6
146	Pure	Ås	16
73	Pure	Ås	17
84, 150	Mixed	Ås	16

After collection the red clover plant-tissue samples were stored at -80^{o}C.

7.5 Extracting and fixating DNA from the Vollebekk greenhouse-individuals

The extraction and fixation of red clover DNA from the plant tissue-samples was performed with a "DNeasy 96 Plant Kit". Since the samples had been stored at -80^{o} C, the instruction "Protocol: Purification of Total DNA from Frozen or Lyophilized Plant Tissue" were used [11]. When the extraction and fixation were performed, only one 96-well-plate was for the red clover tissue samples. The other 96-well-plate were used by someone else. This was for stabilization under centrifugation.

- Some minor modification was done to the protocol.
  - At "stage 26", only 50 µl of AE buffer was added to each red clover sample. Also, the incubation at room temperature (15-25^{o} C) was extended to 5 minutes.
  - At "stage 27" stage 26 was to be repeated. The centrifugation done this time was set to 3800 G-force instead of 6000 rpm.

- To estimate the quality of the DNA extracted, the extracted DNA was put through an agarose gel-electrophoresis (1% agarose).
- A nanodrop was performed to measure the concentration of the DNA extracted. The nanodrop-instrument was calibrated using AE-buffer. The concentrations were found to be in the range 30-70 ng/µl.

7.6 Preparing primer-solutions

The first thing to do is to prepare the primers that is going to be used, if they are not already prepared. The primers were ordered from Thermo Fisher Scientific, and they ship it as a powder stored in small tubes. The primers are needed as a solution, with a concentration of 10 µM. At each tube it says how many nano-moles it contains. I want a solution of 100 µM, so 10 µL milliq-water is added for each nano-mole of primer powder. For example, the tube with the primer TP6_4397690_F3 contains 23.5 nano-moles primer, and since 10 µL should be added for every nano-moles with primer, I add 235 µL, and achieve a primer solution with a concentration of 100 µM. This primer solution is stored in a freezer to later uses. This solution itself is not used as it is but is used to make 10 µM primer-solutions. 10 µL of the 100 µM-solution is mixed with an extra 90 µL of milliq-water, resulting in a 10 µM primer-solution.

Table 17. *Primer pairs tested, with annealingtemperature, and annealingtemperature I ended up with after testing.*

Primer pair	**Annealingtemperature (ºC)**	**Final annealing temperature**
TP6_4397690 F3R3	55	55
ITS	55	55
TP4_23347267 F3R3	55	57
TP4_11383593 F2R2	55	59
TP2_18520944 F1R1	55	55
TP4_11383593 F1R1	55	57
TP2_18520944 F3R3	55	57
TP2_18520944 F4R4	55	57

Table 18. Primer-pairs picked for test-cutting with restriction-enzyme.

Primer-pair	**Restriction-enzyme**	**Temperature for cutting (°C)**
TP6_4397690 F3R3	BsTU1	60
TP4_23347267 F3R3	Sca1	37
TP4_11383593 F1R1	Mse1	37
TP4_23347267 F1R1	Sca1	37
TP2_18520944 F3R3	Hae3	37
TP2_18520944 F4R4	Hae3	37

Table 19. After cut-testing different primer-pairs, these are the primer-pairs that were chosen to be used in genotyping

Primer-pair	**Final annealing temperature (°C)**	**Restriction-enzyme**	**Temperature for cutting (°C)**
TP6_4397690 F3R3	55	BsTU1	60
TP2_18520944 F4R4	57	Hae3	37

Some of the samples didn`t have a clear enough image to make an assessment of the genotype, or a well might have been clear. Both the TP6_4397690 F3R3 and TP2_18520944 had such samples, and they both had a round-up test where all these samples were amplified and cut again.

A vital difference between the test-cut and the genotyping-cut was the recipe for cut-mix. It was the "ordinary" pr. reaction recipe for TP6_4397690 F3R3. TP2_18520944 column 1 and 2 on the other hand had somewhat weak on unclear gel-images with only 0.2 µL, and this was increased to 0.4 µL enzyme for column 3 and 4, and the quantity with PCR-product was increased from 10 µL to 20 µL. The buffer increased from 2 µL to 3 µL. The quantity of milli-Q water was decreases from 7.8 µL to 6.6 µL. All in all, the new reaction volume for the cut-reaction was 30 µL. Before the samples was applied to the gel after being cut, 6

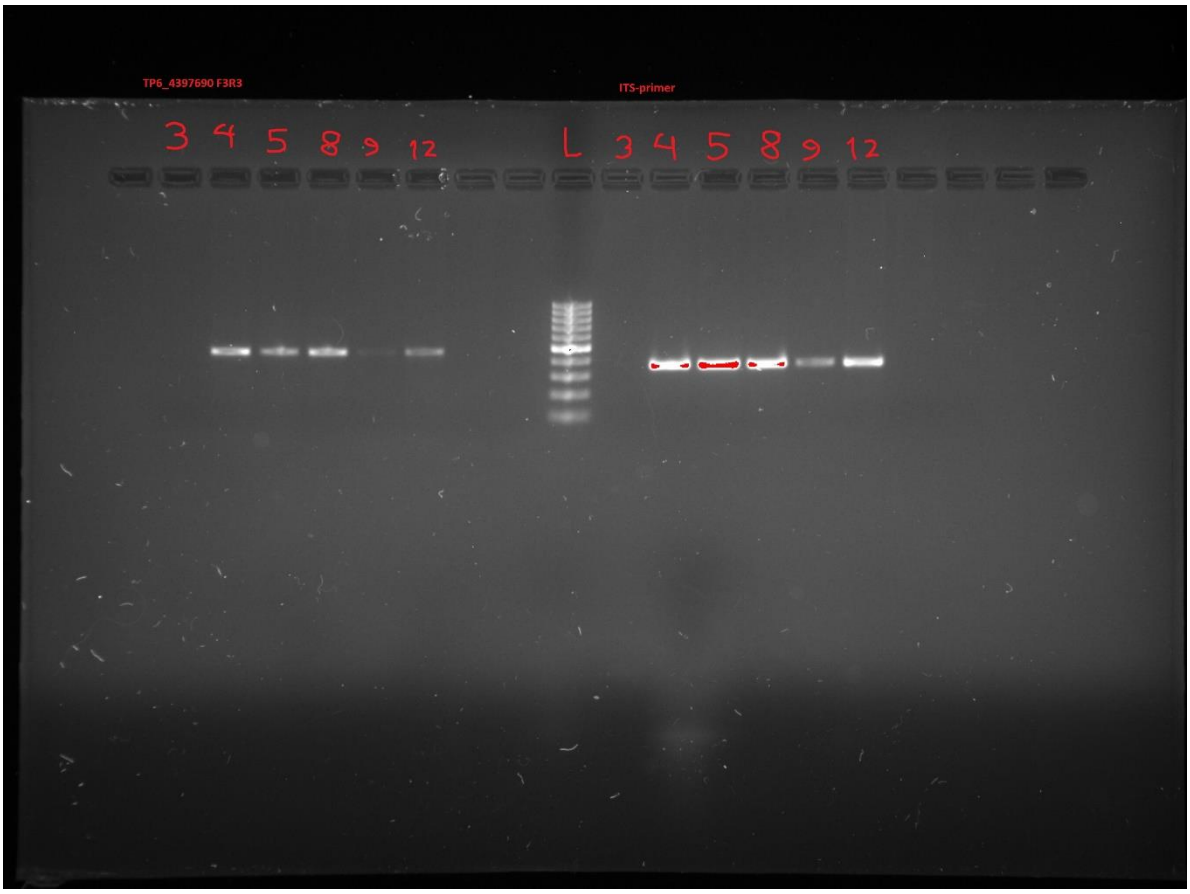
µL loadingbuffer was added to all the samples that were cut (not sample 4 and negative control). The gel-image still seemed a little unclear, so it was decided to increase the amount of restriction-enzyme in the cut-mix further, to 0.8 µL pr. reaction. It`s still just 3 µL buffer pr. reaction, but the amount of milli-Q water pr. reaction is changed to 6.2 µL. 20 µL PCR-product was still used. The cut reaction is now also run at 37^{O}C for 30 minutes. This gave a somewhat clearer gel-image, so it was decided to keep it this way for the rest of the cuts. Columns cut this way were 5-12. Genotyping of samples that showed an unclear or uncertain genotype were repeated. This testing was done with the 0.8 µL enzyme-recipe for the cut-mix.

7.7 Development of CAPS-markers

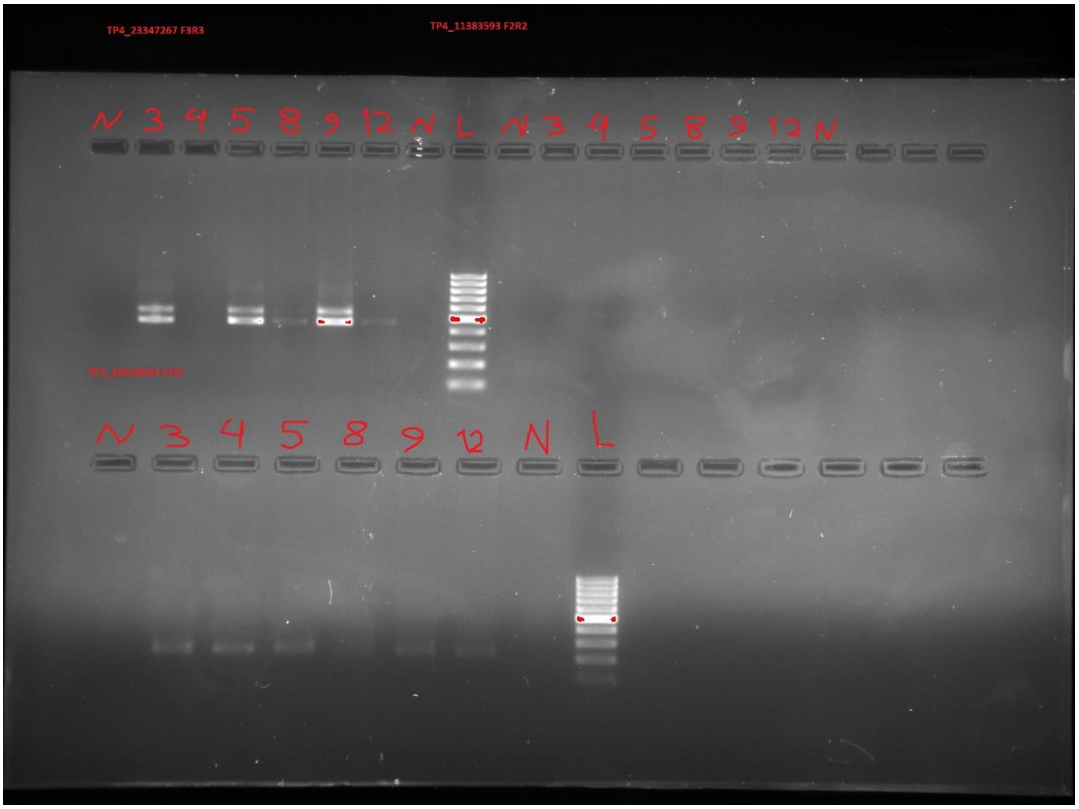
Statistics, by calculation of F_{ST}-values and CHI^2-testing, were used to find which SNPs that appeared with different frequencies in pure-stands and mixed-stands. Based on which SNPs that were deemed to be significant different in pure -and mixed stands an effort were put into finding CAPS-markers that could tell us if an individual was grown in a “pure stand” or “mixed stand”. Possible primers were found for all SNPs deemed to be significant at p-level= 0.1. Only a few of them however were chosen for testing:

- TP6_4397690 F3R3
- TP4_23347207 F3R3
- TP4_11383593 F2R2
- TP2_18520944 F1R1

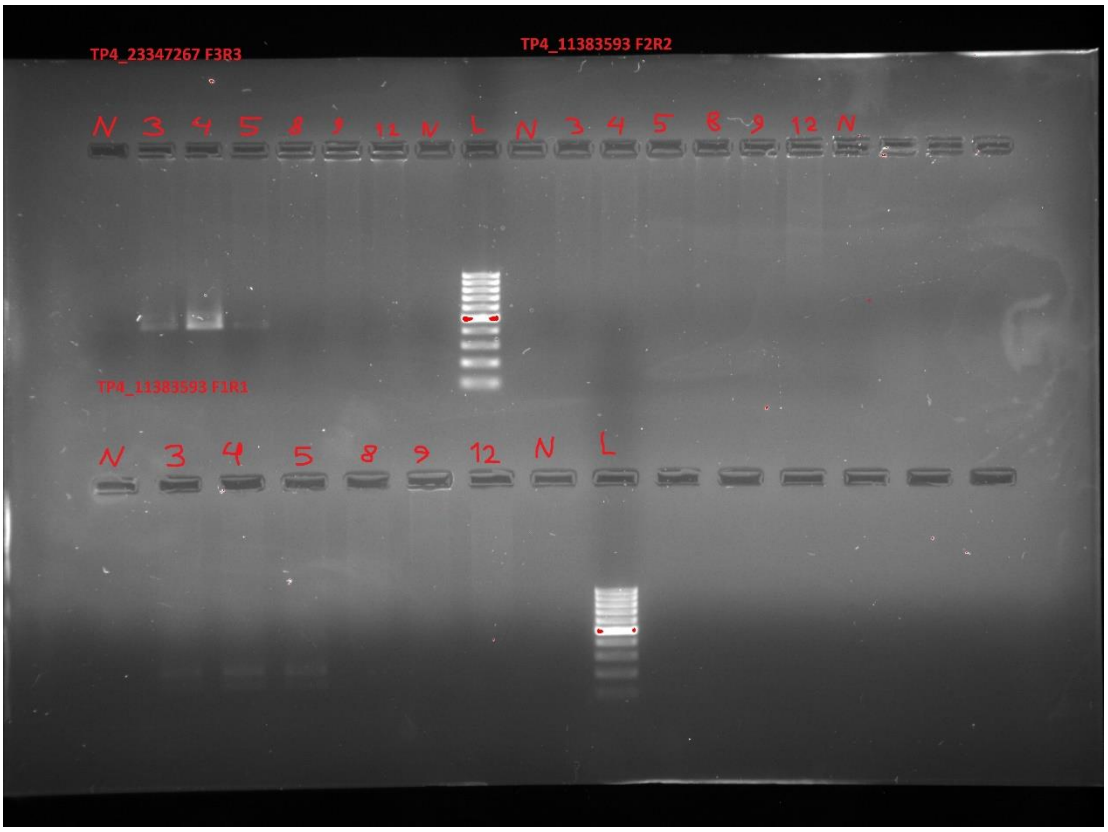
These were tested by first having them PCR amplified, and then having their gel-image studied after the gel-electrophoresis was done. This resulted in the images below:



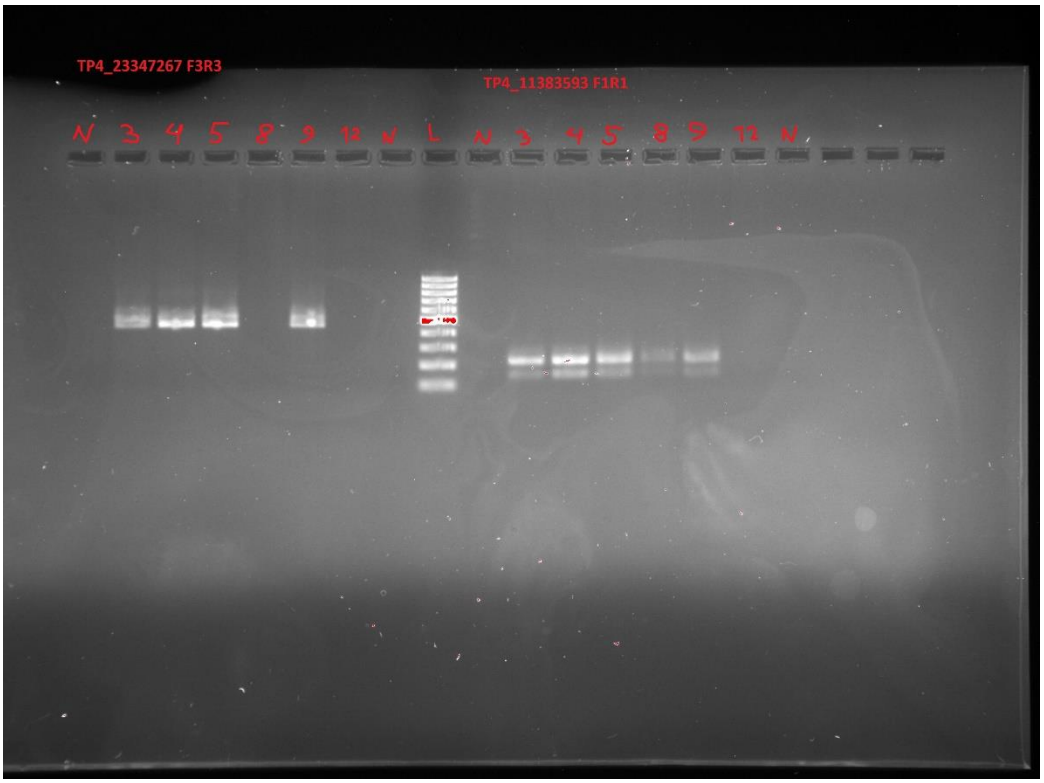
Gel 1. *Gel-electrophoresis of the TP6_4397690 F3R3 -and ITS-primer products*



Gel 2. *Gel-electrophoresis of the TP4_23347267 F3R3 -and TP4_11383593 F2R2 primer products.*



Gel 3. *Gel-electrophoresis of the TP4_23347267 F3R3, TP4_11383593 F2R2 -and TP4_11383593 F1R1 primer products*

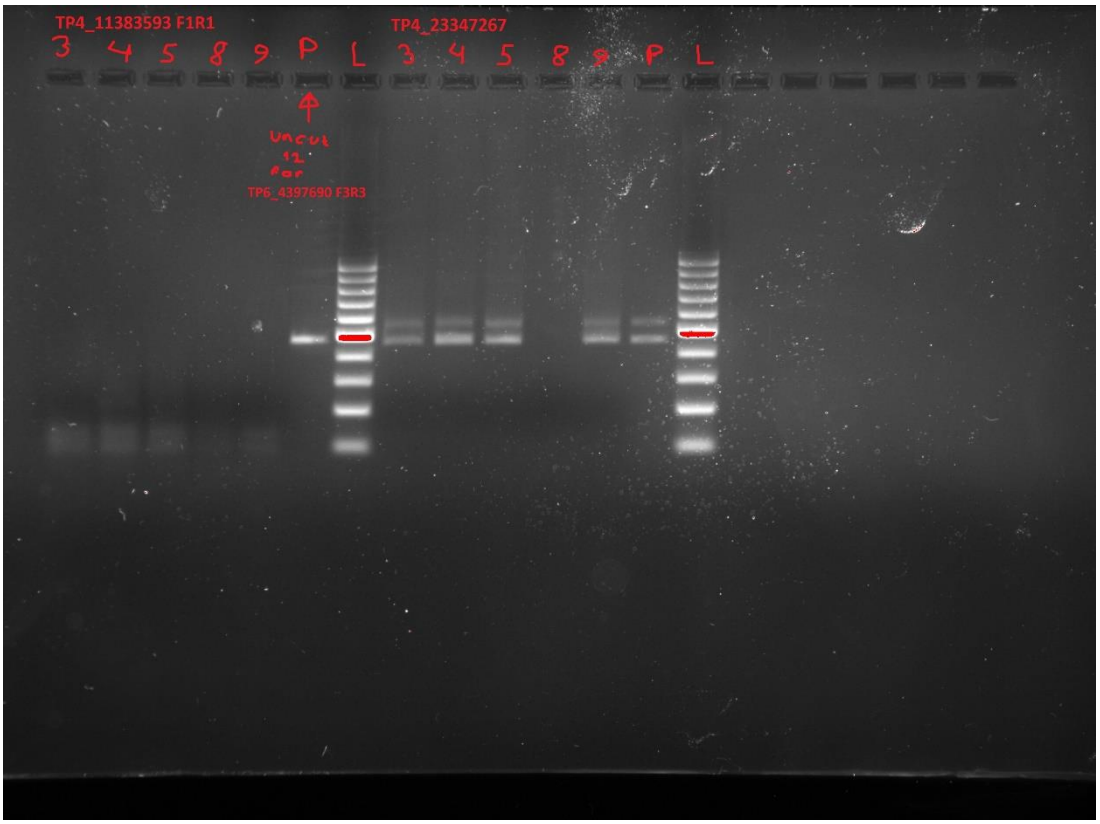


Gel 4. *Gel-electrophoresis of the TP4_23347267 F3R3 -and TP4_11383593 F1R1 PCR products*

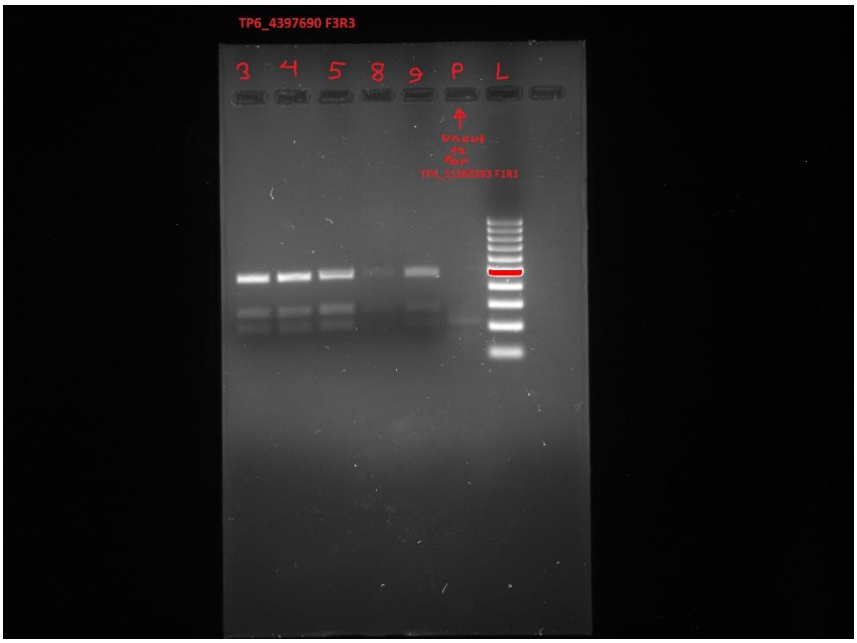
By studying the gel-images, a set of primers were chosen for further testing with restriction-enzymes. These were the primers that showed a clear and strong “PCR-image”. These were tested to see which of them also would have clear image with distinguishable genotypes. The primer products that were test-cut were:

- TP6_4397690 F3R3
- TP4_23347267 F3R3
- TP4_11383593 F1R1
- TP2_18520944 F3R3
- TP2_18520944 F4R4
- TP4_23347267 F1R1

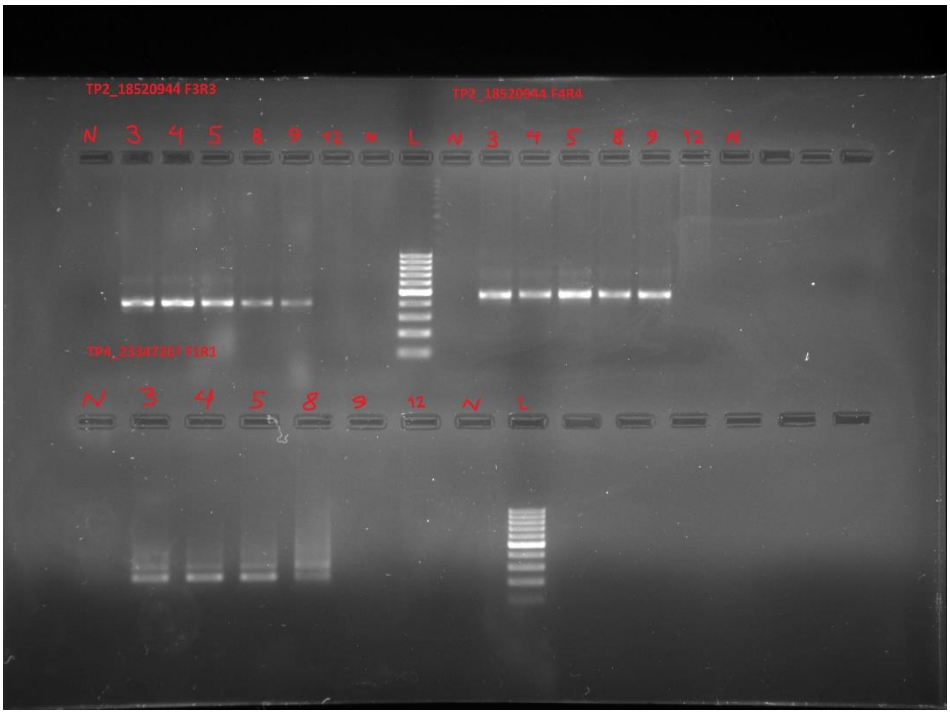
These were test-cut by using PCR-products from earlier we had stored in a freezer. These were cut with appropriate restriction-enzyme, and then went through a gel-electrophoresis.
This resulted in the images below.



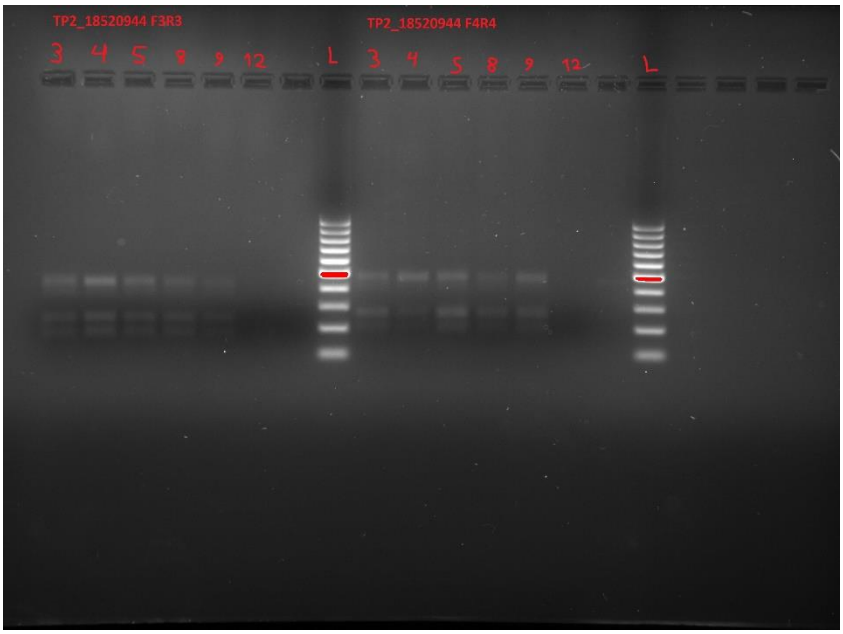
Gel 5. *The gel-electrophoresis of restriction-enzyme digested TP4_11383593 F1R1 -and TP4_23347267 F3R3 PCR-product.*



Gel 6. *The gel-electrophoresis of restriction-enzyme digested TP6_4397690 F3R3 PCR-products.*



Gel 7. *The gel-electrophoresis of restriction-enzyme digested TP2_18520944 F3R3, TP2_18520944 F4R4 -and TP4_23347267 F1R1 PCR-products.*

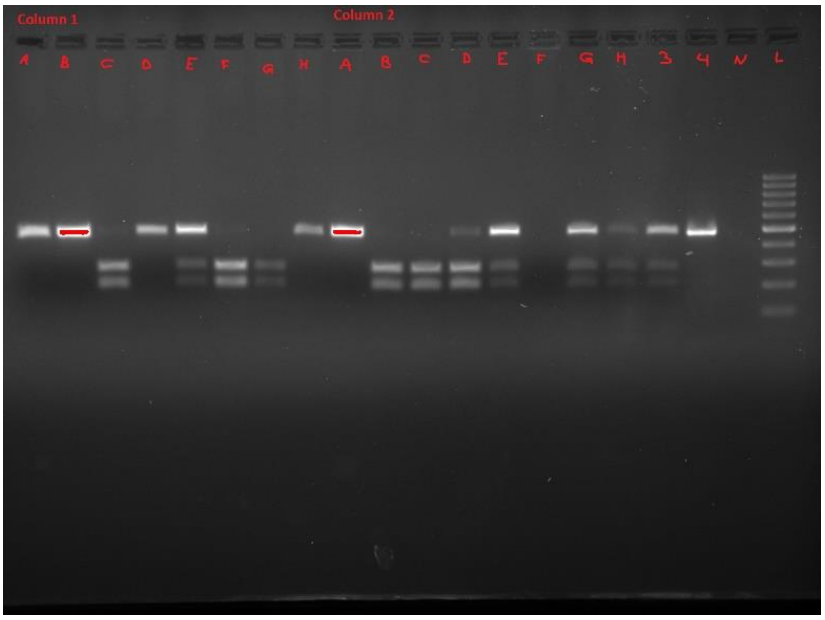


Gel 8. *The gel-electrophoresis of restriction-enzyme digested TP2_18520944 F3R3 -and TP2_18520944 F4R4 PCR-products*.

After the test-cut of the different primer PCR-products were done, two primers were chosen:

- TP6_4397690 F3R3
- TP2_18520944 F4R4

These two primer-pairs, with corresponding restriction-enzymes, were used to genotype the Vollebekk greenhouse individuals. The greenhouse individuals DNA-samples were stored in tubes, eight tubes in a column, and 12 columns in a tray of 96 individuals in total. Since it`s only two different primer-pairs that is used to find the genotype of the greenhouse individuals, it means that in the end it was decided to only genotype two SNPs that are mentioned in the name of the primer-pairs. The digestion of the PCR-products with their corresponding restriction-enzymes is shown in the images below.



Gel 9. *The digestion of column 1 and 2`s TP6_4397690 F3R3 -products.*

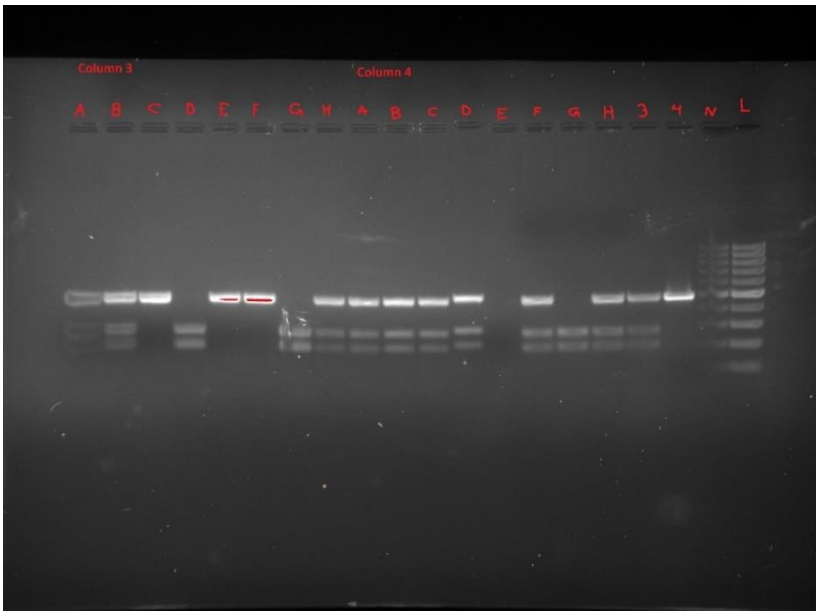
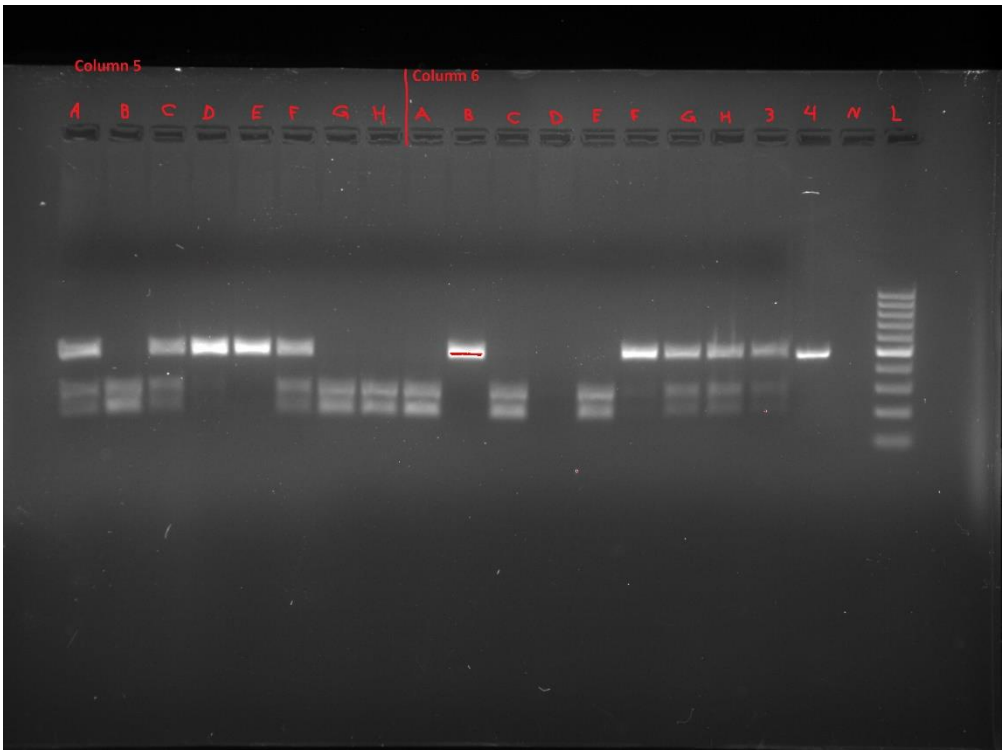
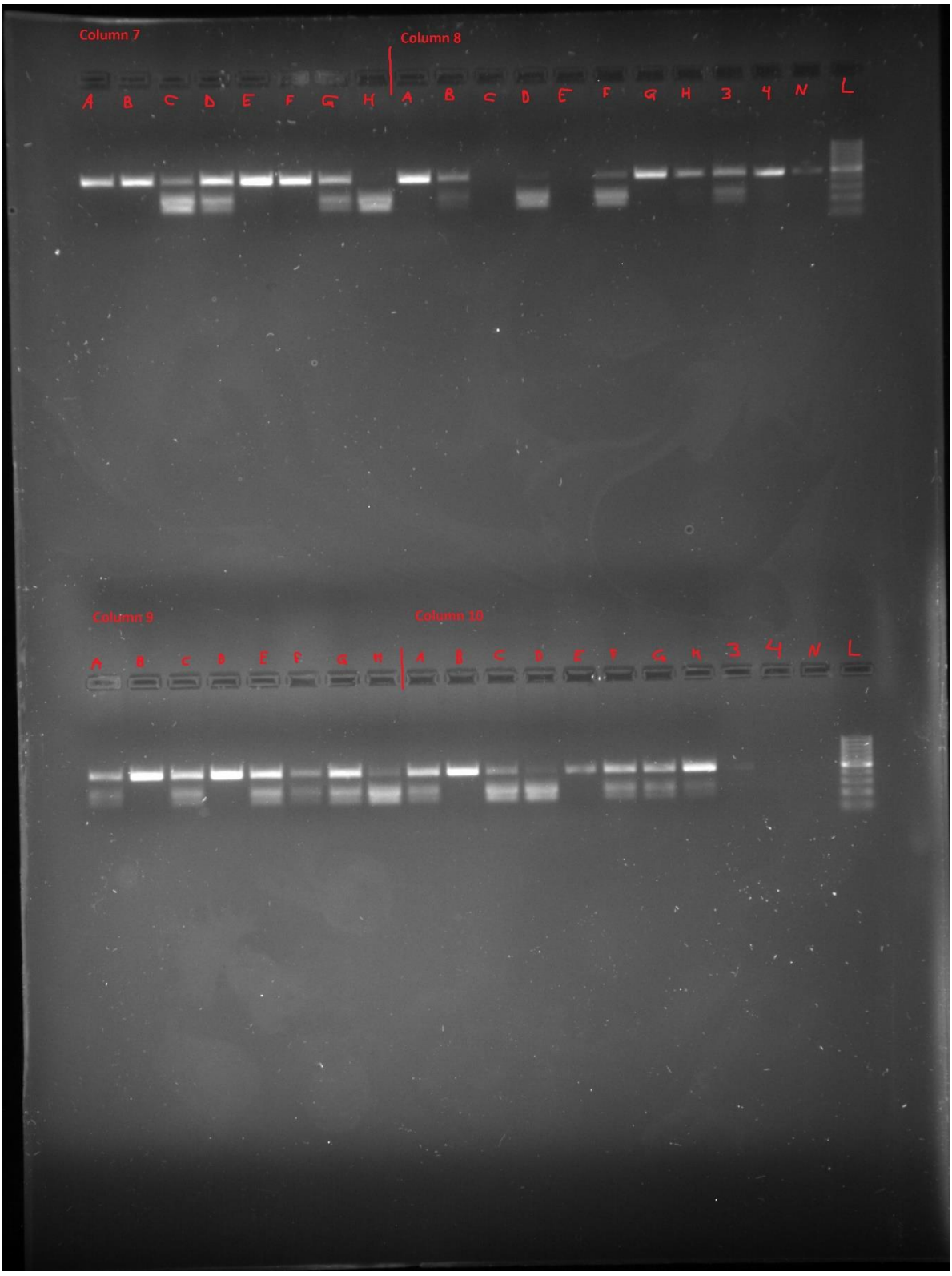


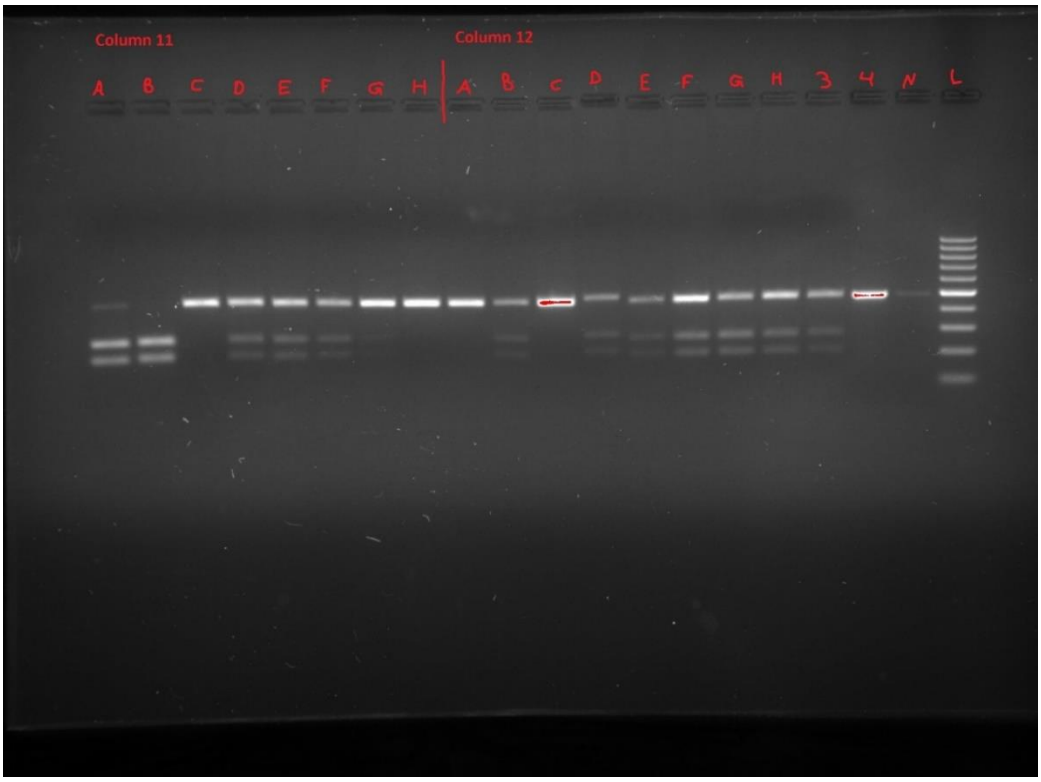
Image 10. *The digestion of column 3 and 4`s TP6_4397690 F3R3 -products.*



Gel 11. *The digestion of column 5 and 6`s TP6_4397690 F3R3 -products.*

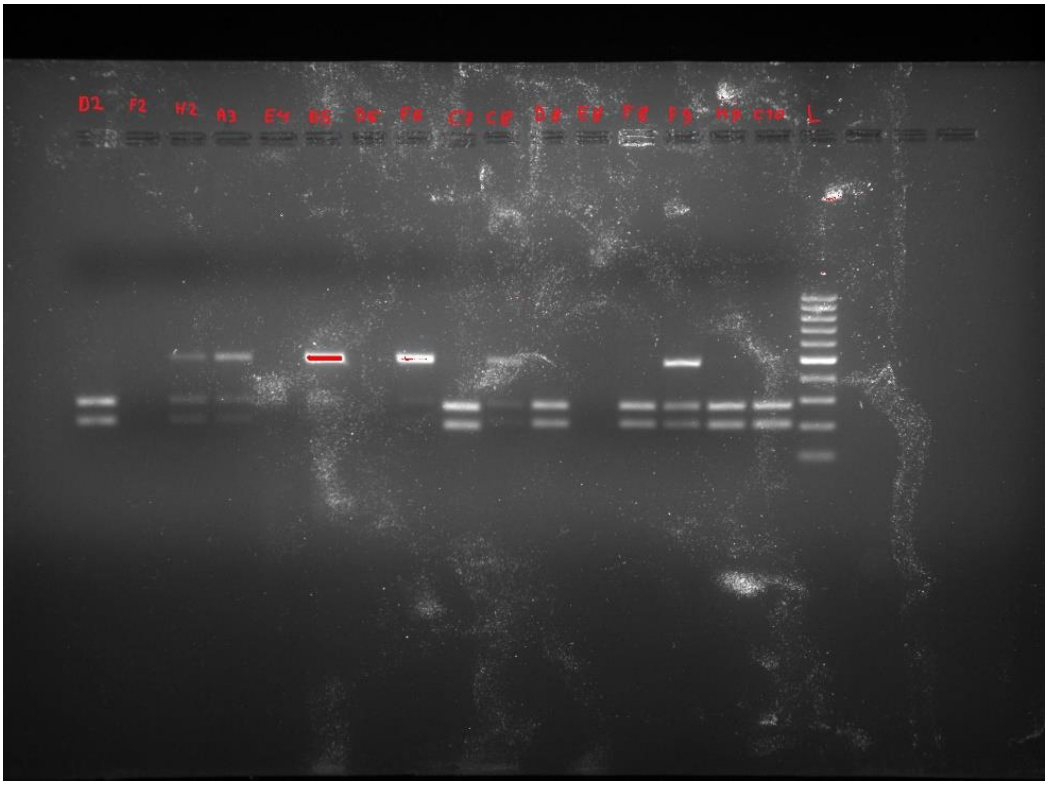


Gel 12. *The digestion of column 7, 8, 9 and 10`s TP6_4397690 F3R3 -products.*

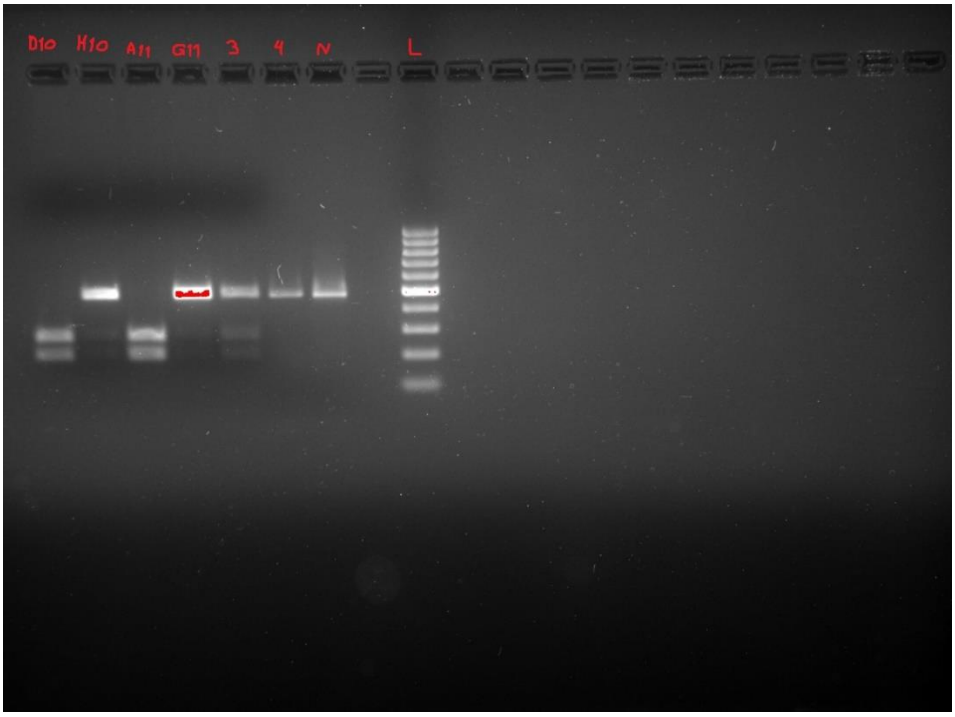


Gel 13. *The digestion of column 11 and 12`s TP6_4397690 F3R3 -products.*

Some wells had an unclear result, or the well might not have shown anything at all (blank). These individuals were tested again.

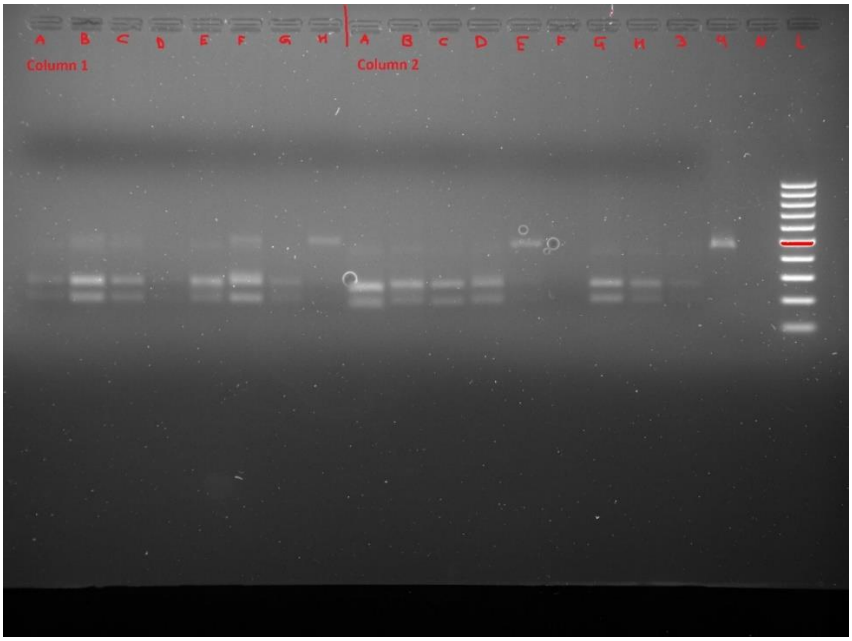


Gel 14. *The digestion of some DNA-samples PCR-products, a second time (TP6_4397690 F3R3).*

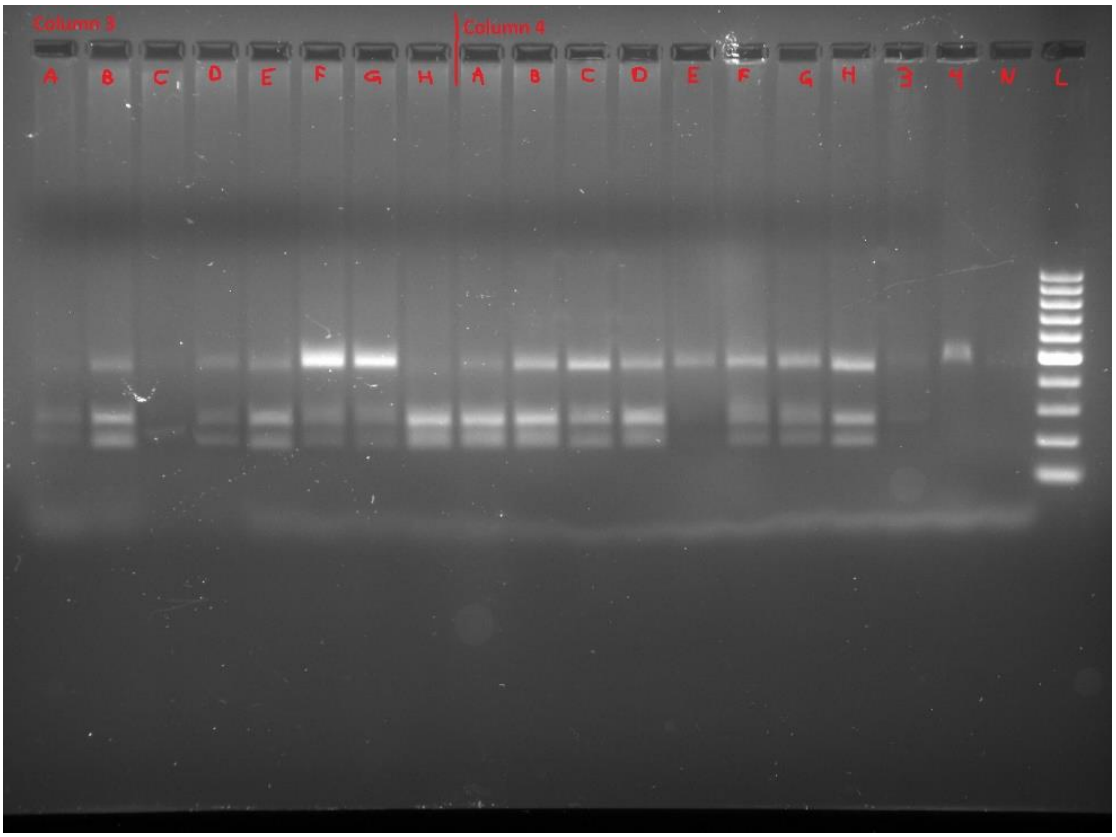


Gel 15. *The digestion of some DNA-samples PCR-products, a second time (TP6_4397690 F3R3).*

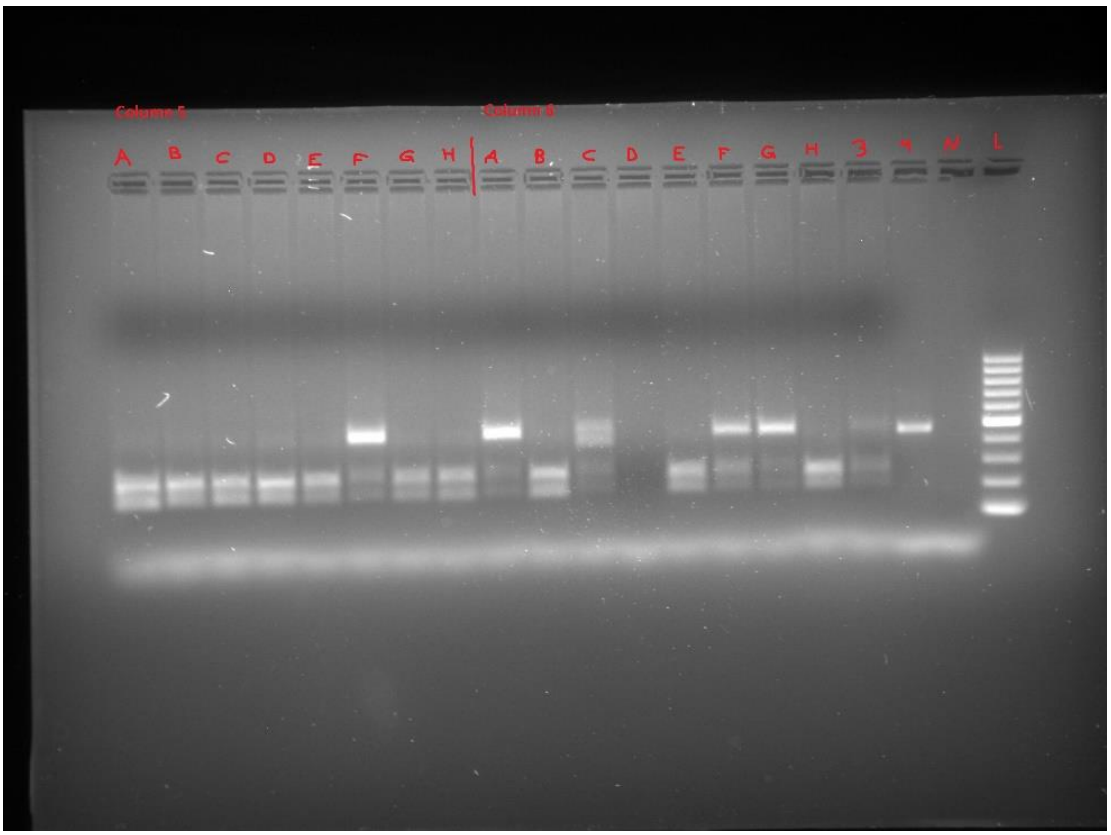
After being done with the digestion of TP6_4397690 PCR-products (with BsTU1), I start to digest the TP2_18520944 F4R4 PCR-products of the greenhouse individuals. Column 5-12 was digested with restriction-enzyme for 30 minutes instead of the ordinary 15 minutes. This was done in hope of getting a more complete digestion and a clearer gel-image. The gel-images is seen below.



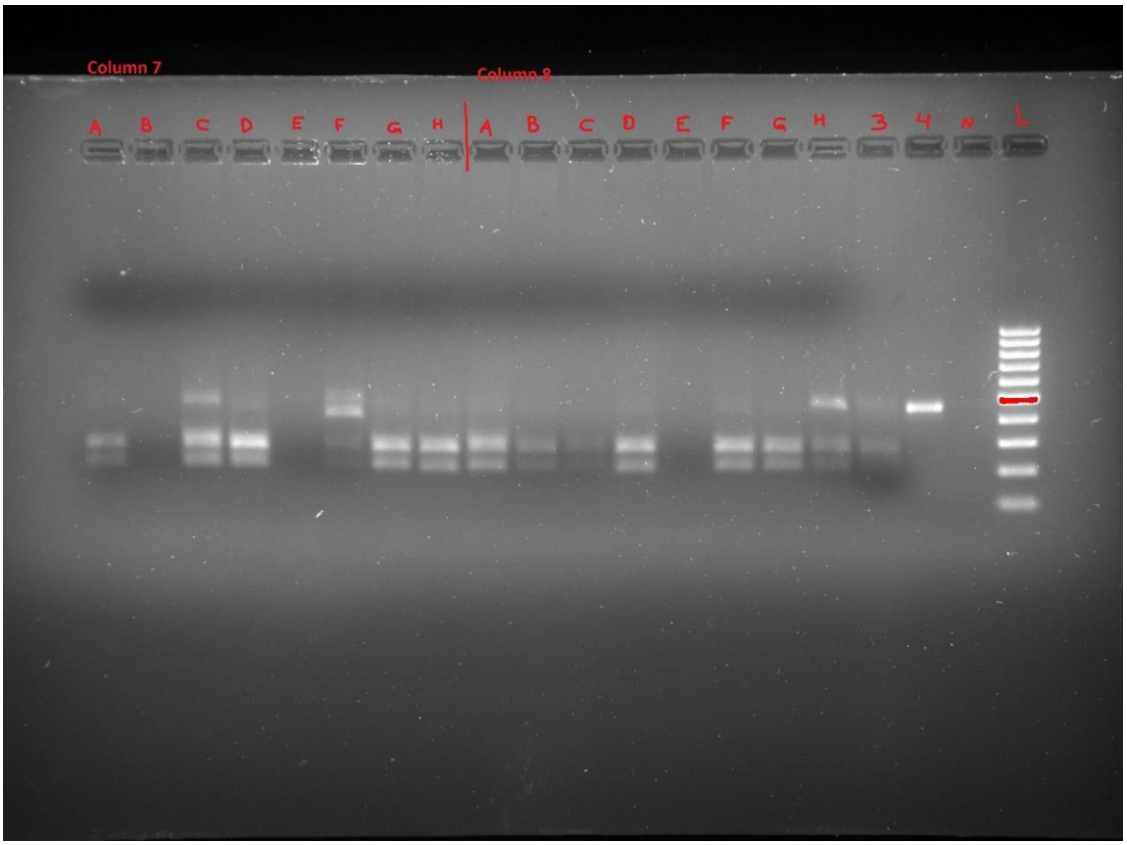
Gel 16. *The digestion of column 1 and 2`s TP2_18520944 F4R4 PCR-product.*



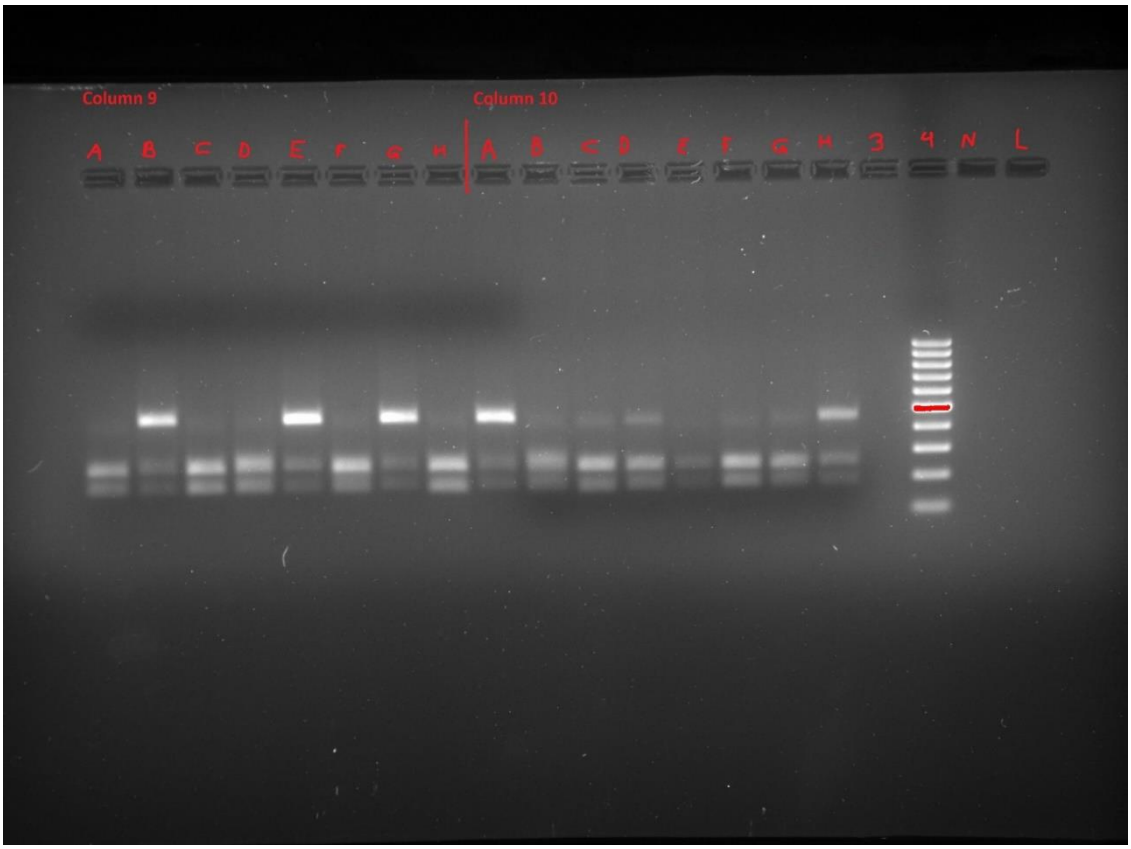
Gel 17. *The digestion of column 3 and 4`s TP2_18520944 F4R4 PCR-product.*



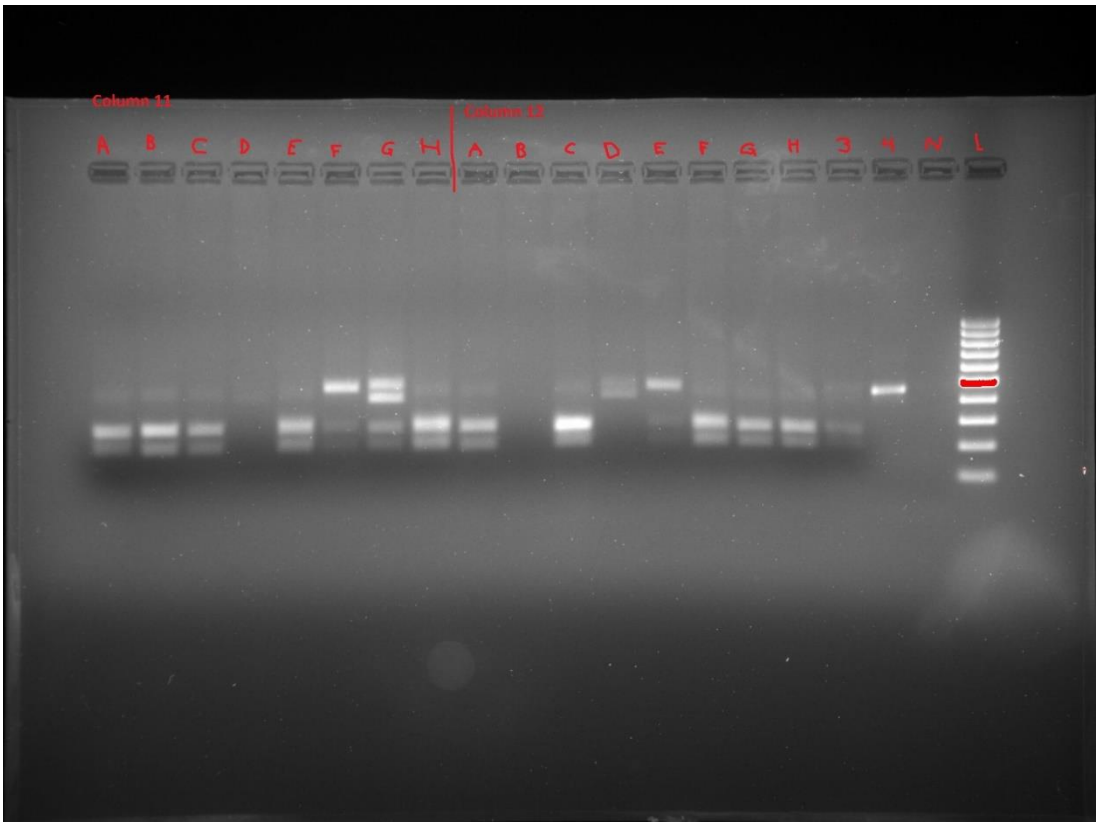
Gel 18. *The digestion of column 5 and 6`s TP2_18520944 F4R4 PCR-product*



Gel 19. *The digestion of column 7 and 8`s TP2_18520944 F4R4 PCR-product*.

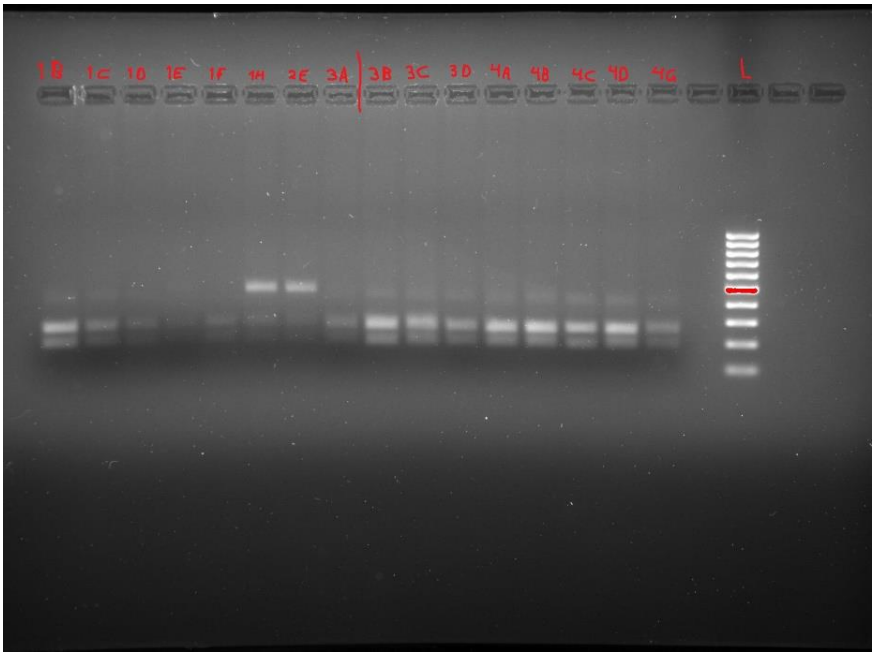


Gel 20. *The digestion of column 9 and 10`s TP2_18520944 F4R4 PCR-product*.

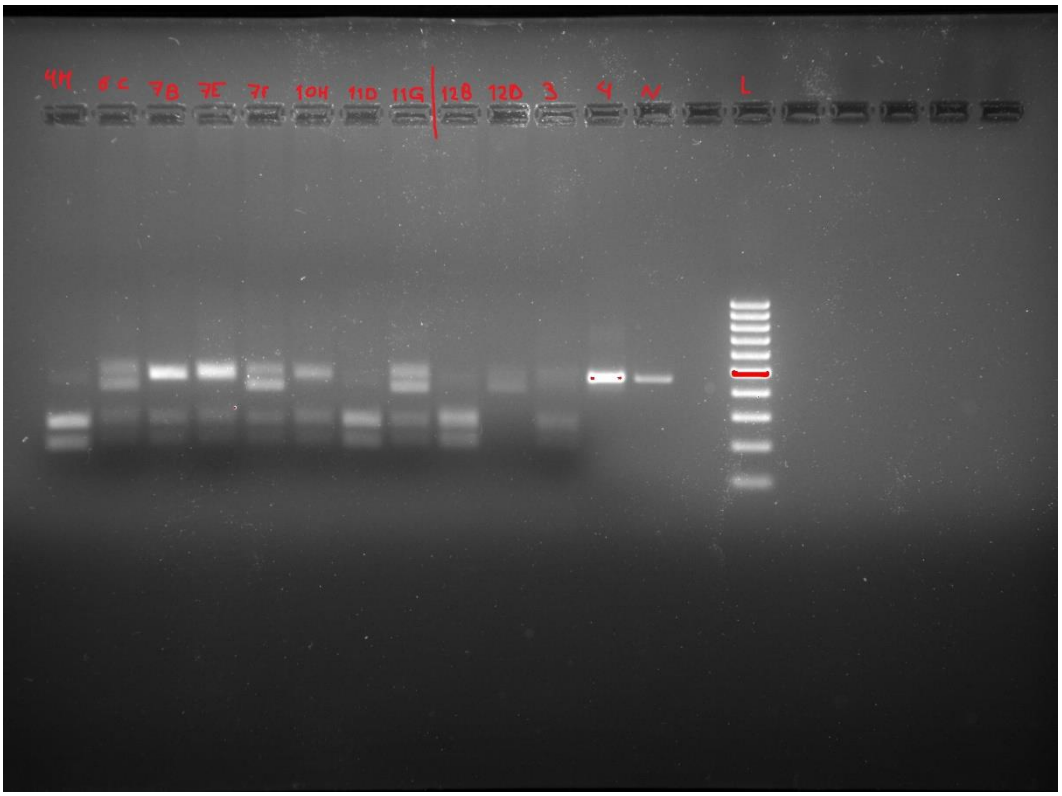


Gel 21. *The digestion of column 11 and 12`s TP2_18520944 F4R4 PCR-product.*

It`s not easy to distinguish which genotype some of the DNA-samples show, so also this time some DNA-samples are “genotyped” again to get a clearer and more distinguishable gel-image.



Gel 22. *The digestion of some DNA-samples PCR-products, a second time (TP2_18520944 F4R4).*



Gel 23. *The digestion of some DNA-samples PCR-products, a second time (TP2_18520944 F4R4).*

Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway