Norges miljø- og biovitenskapelige universitet

Masteroppgave 2016 60 stp
Institutt for plantevitenskap

Identifisering av kromosomområder som kontrollerer tidspunkt for stengelstrekning i rødkløver med bruk av Genotyping by Sequencing (GBS)

Identification of regions that control timing of elongation in red clover using Genotyping by Sequencing (GBS)

Øystein Westerhagen Milvang
Plantebioteknologi

Forord

Oppgaven er en avslutning på masterstudiet i plantevitenskap, med fordypning i plantebioteknologi, ved Norges miljø- biovitenskapelige universitet på Ås. Det har vært en spennende og lærerik periode som jeg er veldig takknemlig for. Feltarbeidet foregikk på Vollebekk forsøksgård i Ås og laboratoriearbeidet har blitt utført ved Cigene på Ås. Sekvensering ble utført eksternt på Cornell University.

Jeg vil takke min veileder Åshild Ergon som alltid har vært tilgjengelig for hjelp og støtte gjennom hele prosessen.

Takk til Øyvind Jørgensen for hjelp i drivhuset på Vollebekk forsøksgård.

Takk til Anne-Guri Marøy og Sylvia Sagen Johnsen fra Cigene som lærte bort og hjalp meg med laboratoriearbeid. Takk til Agropro somfinansierte oppgaven.

Jeg vil og takke min far Otto Milvang som kodet et skript essensielt for filtrering av data.

Takk til Mallikarjuna Rao Kovi for gode råd og skript som hentet ut sekvenser fra datafil.

Og takk til min kjære Olivia og sønn Daniel som har støttet meg og vist tålmodighet når jeg har vært fraværende på grunn av arbeid med masteroppgaven.

Ås, Desember 2016

Øystein Westerhagen Milvang

Sammendrag

En populasjon rødkløver av (*Trifolium pratense* L., sorten «Lea»), ble karakterisert for tidspunkt for begynnende stengelstrekning (dager til strekning – DTS). Dette er det første synlige steget i overgangen fra vegetativ til blomstrende fase. Av de 672 plantene som ble testet ble 52 av de tidligste (gjennomsnittlig DTS på 37,7) og 52 seneste av de seneste (gjennomsnittlig DTS på 80,5) valgt ut til genetisk karakterisering. Seks uker etter gjennomsnittlig tidspunkt for stengelstrekning i hver av gruppene hadde den tidlige gruppen dobbelt så høy andel skudd i strekning. Genotyping by sequencing (GBS) er et kraftig verktøy for å plukke opp genetisk variasjon. Vanligvis brukes GBS på individnivå men er her blitt brukt på DNA-pooler og antall reads ble brukt til å anslå allelfrekvenser. De 52 individene i hver gruppe ble fordelt på tre tilfeldige undergrupper, og lik mengde DNA fra hvert individ i undergruppen ble kombinert i en «pool» så det ble totalt seks «pooler». SNPer med signifikant forskjellig allelfrekvens i tidlig og sen gruppe ble funnet ved å regne ut parvise Fst-verdier for alle undergrupper mot gjennomsnittlig allelfrekvensen for undergruppene av den andre genotypen (totalt seks fst verdier), og sjekke størrelsen av disse. Sekvensering ble gjort to ganger med ulike restriksjonsenzymer, Pst1 og Apek1. Med Apek1 ble det funnet flest SNPer men kvaliteten av SNPene var bedre blant de funnet med Pst1. Totalt fant jeg 63 SNPer med signifikant forskjellig allelfrekvens mellom den tidlige og sene gruppen ($P<0,01$ i alle seks fst sammenligninger, false discovery rate <0,001). Signifikante SNPer ble funnet i alle kromosomer med høyest forekomst i kromosom 6 og 7. Bare èn SNP ble funnet med begge enzymer. Noen av SNPene som ble funnet så ut til å ligge i nærheten av QTL for blomstringstid og vinteroverlevelse i som tidligere har vært identifisert i andre populasjoner av rødkløver samt i område som er synteniske til områder i *Medicago truncatula* med QTL for blomstringstidspunkt. To SNPer som ble funnet kan styrke muligheten for at CONSTANS osom kontrollerer blomstringstid i mange andre planter har homolog som spiller en rolle i tidspunkt for strekning i rødkløver.

Abstract

A population of red clover (*Trifolium pratense* L., cultivar «Lea»), was characterized for timing of stem elongation (Days to elongation – DTS). This is the first visible sign of transition from vegetative to flowering phase. Of the 672 plants tested, 52 of the earliest (average DTS of 37,7) and 52 of the latest (average DTS of 80,5) were chosen for genetic characterization. Six weeks after average time of elongation within the two groups the early elongating group had twice the proportion of elongated shoots. Genotyping by sequencing (GBS) is a powerful tool to detect genetic variation. Commonly GBS is performed on individuals but is here performed on DNA-pools and the number of reads is used to estimate allele frequency. The 52 individuals from each group were randomly divided into three subgroups and equal amount of DNA from each individual in each subgroup was combined in a "pool", creating a total of six "pools". SNPs with allele frequency that differ significantly between early, and late plants was found by calculating pairwise Fst-values for all subgroups against the average allele frequency of the subgroups of the other phenotypes (a total of six Fst values), and checking the value of these. Sequencing was performed twice, using different restriction enzymes, Pst1 and Apek1. The GBS-run with Apek1 detected most SNPs but the overall quality of the SNPs from Pst1 was better. I found a total of 63 SNPs with significantly different allele frequency between the early and late group ($P<0,01$ in all six Fst comparisons, false discovery rate <0,001). Significant SNPs was found in all chromosomes with the highest density in chromosome 6 and 7. A single significant SNP was detected with both enzymes. Some of the SNPs appeared to be located near QTLs for time of flowering and winter-survival found in other populations of red clover. Also some SNPs appeared to be located in areas syntenic to QTL for time of flowering in *Medicago truncatula*. One of the findings suggests that *CONSTANS,* a gene that controls time of flowering in other plants have a homologue in red clover that plays a role in the timing of flowering.

Innholdsfortegnelse

1. Introduksjon

Rødkløver (*Trifolium pratense* L.) er en av våre viktigste kulturplanter. Den er nitrogenfikserende ved hjelp av symbiose med bakterien *Rhizobium leguminosarum* bv. trifolii.. Det sies at planten var med på å tilrettelegge for den industrielle revolusjon ved å føre nitrogen tilbake til jorda, som hadde blitt tapt ved transport til byene. Bruken av rødkløver gikk drastisk ned etter bruken av kunstgjødsel ble dominerende i landbruket, muliggjort av Haber-Bosch metoden for produksjon av ammonium fra atmosfærisk nitrogen. Rødkløver er fremdeles en viktig plante i økologisk drift og bør kanskje tas mer i bruk for å gjøre jordbruket mer bærekraftig ved å redusere behovet for nitrogengjødsel. Overforbruk av nitrogengjødsel fører til eutrofiering og utslipp av lystgass, N_2O som bidrar til global oppvarming og sur nedbør. Rødkløver er, og har lenge vært vår viktigste belgvekst i Norge fordi godt egnet til grønngjødsling og i fôr. Sortene vi bruker i Norge er godt tilpasset klimaet etter kultivering i minst 200 år og trolig mye innførsel av hardføre frø av utenlandsk opprinnelse (Marum/Skog og Landskap). Planten er flerårig og lever vanligvis i 2-4 år i dyrka eng (Christie & Martin 1999). Første år gir som oftest dårligere avling enn påfølgende år. Det er på grunn av etableringstid er avlingen ikke alltid god første år og eventuell nedgang i årene etter er assosiert med dårlig varighet (Barnhart & Rueber 2013). Planten vokser som tuer med flere skudd. Ved blomstring avtar den vegetative veksten og stenglene strekker seg. Planten har kraftig vekst og tåler flere høstinger i året. Generativ vekst forbindes med nedgang i ny biomasseproduksjon, men mengden tørrstoff øker og fordøyeligheten reduseres (Buxton et al. 1985; Cassida et al. 1999). Det er det vanlig å slå enga når de fleste har startet stengelstrekning og omtrent 20% av plantene blomstrer (Undersander et al. 1990; Wiersma & Bolen 2000). Hvordan tidspunktet for stengelstrekning i rødkløver kontrolleres genetisk ville vært svært nyttig å ha mer kunnskap om for videre foredling. Rødkløver som strekker seg tidlig vil ha en fordel i konkurranse om lys. Mesteparten av studier samt frøproduksjon på rødkløver er utformet slik at kløveren ikke konkurrerer med andre arter. Et nytt studie foreslår at det i blandede kulturer er mer seleksjon for tidligere strekning enn i rene kulturer (Ergon & Bakken 2016). De fleste varianter rødkløver er ikke tilpasset det kalde klimaet i nord, korte vekstsesong og lange dager. Derfor er det fremdeles stort potensiale for foredling for bedre tilpasning til våre breddegrader (Helgadoˊttir et al. 2000). Sen-blomstrende varianter av rødkløver er knyttet til bedre vinteroverlevelse (Choo 1984). Vinteroverlevelse er også knyttet til mengden adventivrøtter fra kronen (Therrien & Smith 1960), og det er vist at det kan være en sammenheng mellom adventivrøtter og blomstringsmønster (Montpetit & Coulman 1991).

Rødkløveren er en langdagsplante med daglengdekrav på minst 12 timer (Vince-Prue 1975). Lengre dager fremskynder blomstring ytterligere (Ergon et al. 2016). Daglengde og temperatur er to faktorer som kontrollerer blomstring. Det er vist at vernalisering kan fremskynde blomstring (Lunnan 1989). I et

forsøk med 2 middels tidlige sorter blomstret alle individer ved 16 timer hvorav flesteparten ved 14 timer (S.R. Bowley et al. 1987). Temperatur påvirker dager til stengelstrekning. Temperaturkraver ser ut til å bli mettet over 14°C (Ergon et al. 2016). Det er kjent at vekst og utviklingstid er miljøavhengig og kan variere ved ulik daglengde mellom sorter (Lunnan 1989; S.R. Bowley et al. 1987). Stengelstrekning og blomsterinitiering er koblet men ser ut til en viss grad å styres separat. Det er et lavere daglengdekrav for stengelstrekning enn for blomsterinitiering (Jones 1974).

Fotoperiodisme-signalveien i langdagsplanten *Arabidopsis thaliana* ligger i bladene og involverer et transportabelt signalstoff antatt å være FT-protein som blant annet er kjent for å aktivere blomstring i apikalt meristem. FT-mRNA produseres i bladene og transporteres til apikalt meristem via floem som respons på CO-protein akkumulering under lange dager. CO-protein er regulert av klokkegener slik at CO-mRNA kun akkumuleres sent på dagen. Funksjonen av disse genene er konservert også i belgplantefamilien (Hecht et al. 2005). Det er identifisert totalt 1424 gener involvert i blomsterutvikling hos rødkløver basert på 430bp lange fragmenter av rødkløver testet mot et simulert dataset av *Medicago truncatula*. Genene ble identifisert ved hjelp av BlastP mot flere proteindatabaser og kategorisert med plant GO slim (J. IŠTVÁNEK et al. 2014).

Rødkløver er diploid (2n) med 14 somatiske kromosomer (7x) (Britten 1963), genomstørrelsen er anslått til 0,427 1cx (haploid) pg, det vil si 417,6 mb (Vižintin et al. 2005). Ett picogram tilsvarer 978 megabaser (Doležel et al. 2003). Genomstørrelsen er tidligere anslått til ~440Mb basert på analyse av 2 ulike sorter rødkløver anslått til 436 Mb og 446 mb (Sato et al. 2005).

Next Generation Sequencing er en fellesbetegnelse for moderne teknologier for sekvensering som er utviklet etter første generasjons Sanger sekvensering. Disse nye metodene kan produsere store mengder data på en kostnadseffektivt måte (Metzker 2010). Ved hjelp av enn av disse nye metodene, Genotyping by sequencing (GBS), kan man oppdage store mengder SNPer på en kostnadseffektiv måte (Elshire et al. 2011; Sonah et al. 2013). GBS gjør det mulig med GWAS (Genome-wide association studies) på grunn av at høy markørtetthet kan gjøre det mulig å finne assosierte polymorfismer til et trekk som er i LD med en eller flere av de andre markørene (Glaubitz et al. 2013; Rafalski 2010). Ved GBS av poolet DNA kan allelfrekvenser estimeres med god nøyaktighet(Stephen Byrne et al.). Tassel-GBS er en mye brukt bioinformatikk-pipeline som genererer SNPer fra rå GBS data. Sekvensering utføres med Illumina sekvenseringsteknologi. Sekvensering av hele genomet er mye mer kostbart og er ofte ikke nødvendig. Restriksjonsenzymenene (RE) kutter i alle områder av genomet. Enkelte RE kan til en viss grad unngå repetitive områder, for eksempel Apek1 som er delvis metyleringssensitiv (Elshire et al. 2011). Det må brukes restriksjonsenzymer som lager et overheng på mer enn 1 bp fordi dette

overhenget brukes videre til å ligere på DNA-adaptorer (Elshire et al. 2011). Det benyttes et barcoding system basert på at det hver brønn i 96 platen (eller opptil 386) tilsettes to ulike DNA adaptorer som ligerer til endene av DNA-fragmentene etter fordøying med restriksjonsenzym. Den ene med unik 4-8bp barcode for identifikasjon i 3'-enden og ett overheng komplimentert til det spesifikke restriksjonsenzymet som skal benyttes på 5'-enden. Den andre type adaptoren er lik i alle 96 brønner og festes også med overhenget laget av restriksjonsenzymet. To restriksjonsenzymer som er vanlige å bruke, ApeKI og PstI, har henholdsvis 5 og 6 basepar gjenkjenningssete. Antallet fragmenter bestemmes av lengden på gjenkjennelsessetet på basekutteren og velges for å variere kompleksiteten. ApeKI produserer flere fragmenter enn PstI. ApeKI lager et overheng på 3bp. Gjenkjennelsessekvensen på ApekI er degenerativ og er GCWGC der W kan være A eller T. (Elshire et al. 2011). Den RE fordøyde DNA-sekvensen kalles «insert». Etter ligering av adaptorer i endene blandes DNA fra alle brønnene i en DNA-pool og det tilsettes to type primere. Primer 1 binder til 3' til barcode adaptoren og har en 5' som er komplementær til «flowcell oligo1». Primer 2 binder til 3' av fellesadaptoren med 5' komplimentær til «flowcell oligo 2». Med ligerte primere utenpå adaptorene med DNA i midten har vi et «GBS library». Dette anrikes med PCR reaksjon som favoriserer amplifikasjon av kortere fragmenter. «Tag» er en unik sekvens (minus barcode) opptil 64bp, fra en eller flere «good barcode reads». Tags fra kortere «inserts» enn 64bp er vanlig og blir amplifisert i høyere grad under PCR reaksjonen enn lange fragmenter. Kortere fragmenter oppstår dersom flere RE-sete finnes innenfor 64bp. Antallet ganger en tagsekvens er observert er indikator på kvalitet (istedenfor å gå ut ifra avstand fra starten av sekvensen). På Illumina er vanligvis ellers en negativ korrelasjon mellom kvalitetsscoren og posisjonen i en read. Viderere alignes sekvensene og det detekteres SNPer mellom prøvene som kan skilles ved hjelp av Barcode-adaptoren. Ulemper ved GBS kan være lav dekning. Mange imputeringsmetoder kan brukes men det er ofte ikke nødvendig for estimering av allelfrekvens (Glaubitz et al. 2013). Det er vanlig å bruke GBS på enkelt individer men det er mulig å bruke DNA-pooler. Det er mulig med GBS å presist estimeres allelfrekvens av SNPer over hele genomet ved hjelp av replikate DNA pools av ulike individer ved å gå ut ifra frekvensen av antall reads (Stephen Byrne et al.). En av grunnene til å poole prøvene og spre de over flere brønner er at ved å ha mange replikater vil det styrke påliteligheten ved beregning av allelfrekvens. Ved å ha tre pooler istedenfor en i hver gruppe, tidlig eller sent tidspunkt for stengelstrekning er det mulig å teste hver enkelt pool mot hele den andre gruppen og redusere antall falske positive betraktelig.

2. Materiale og metode

2.1 Materialer og vekstbetingelser

Sorten som ble brukt i forsøket var «Lea», en diploid, middels tidlig sort foredlet av Bioforsk Øst løken, godkjent i 2002 og eid av Graminor. Vinteroverlevelse i sorten er regnet som dårlig til middels god (Graminor). Lea stammer fra de tre sortene Nordi, LGRk8801 og Bjursele og er i dag kanskje regnet som den viktigste sorten i Norge (Marum/Skog og Landskap). Forsøket sammenlignet 672 planter som stammer fra samme frøpakke som ble karakterisert for koblings-ulikevekt av De Vega et al. (2015) (DeVega et al. 2015). Plantene ble kultivert i drivhus på Vollebekk forsøksgård, NMBU. Frøene ble sådd 1. september og det ble veid opp 2,56 g frø (1000-frøvekt ~2 g) fordelt på 3 bakker med Tjerbo gartnerjord (LOG) dekket med et tynt lag sand. Fjorten dager senere ble plantene priklet ut i firkantpotter 105 x 105 x 75 mm (PF310 VEFI, faktisk volum 550 cm^3). Pottene ble jevnt fordelt på 7 bord med 96 på hvert bord med en plante i hver potte. Det ble brukt samme jordtype, men da uten sandlaget på toppen. Det sto 8 planter i hvert pottebrett som hadde plass til 15 potter så et var tomme plasser imellom. Daglengden var satt til 20 timer, 14 stk HQI/HPI-T lamper med omtrent 200 µmol-$1m^2s$ som grense for å skru på tilleggs-lyset. Temperaturen var regulert til så nær 16°C som mulig. Plantene ble gjødslet fra uke 5, en gang i uken med gjødslingsvann. Dette hadde en ledningsevne på 1,5 ms og var basert på en blanding av vann og 50/50 av 2 stamløsninger: 60 g/l Kalksalpeter (YARA, Calcinit 15,5N + 19Ca) og 80 g/l (YARA, Kristalon 9-11-30 + 7MgO + Mikro). Stamløsningene hadde begge ledningsevne på 14,8ms. Lysstyrken på tilleggs-lyset ble målt kvelden 9.10.15 med APOGEE MQ-200 Quantum meter. Måling ble gjort ved 30cm høyde over bordet på topp, midt og bunn av de 7 bordene. Det var stor forskjell i lysstyrke, gjennomsnittlig lysstyrke var 92 $\mu mol^{-1}m^2s$. Med standardavvik på 26,3 mellom bordene (Vedlegg 1 Tabell 4)

2.2 Registreringer i drivhus

Antall dager til begynnende stengelstrekning etter såing(DTS) ble registrert for hver enkelt plante. Registrering av hvilke planter som hadde startet å strekke seg ble gjort 18 ganger fra 22.10.15 frem til 4.12.15, med unntak av den aller første registreringsdatoen da jeg ikke hadde tid til å se gjennom alle (Vedlegg 1 tabell 2). Starttidspunkt ble definert som stengelstreking på minst 2 cm. Dato og lengden fra base til ytterste node ble notert. Hvis strekningen ble oppdaget sent og stengelen var lenger enn 2 cm ved registrering ble lengden over 2 cm dividert en verdi som representerte hastigheten av strekningen som cm/dag. Registreringsdatoen ble trukket fra denne verdien for å estimere DTS når en plante ble oppdaget flere dager etter strekningen var forbi 2 cm. For korrigerte verdier ble tallet rundet av til

nærmeste hele tall.

DTS = RegDato – (Lengde - 2 cm * EstDagligVekst) – SåDato

Stengelstrekning på gjennomsnittlig 1,17 cm per dag med standardavvik på 0,4 ble anslått ved at 12. november ble 32 planter som allerede var observert i strekning målt på nytt, 2-17 dager etter estimert DTS (Vedlegg 1, tabell 3). Første registrering av strekning ble gjort 22-23 oktober, det vil si 52 dager etter såing. Det ble allerede da registrert i alt 164 planter i strekning. De 52 tidligste og 52 seneste plantene i strekning ble valgt ut videre fenotypisk og genotypisk karakterisering. Planen var egentlig å bruke tre pooler à 20 individer til hver av de to gruppene. Antallet ble minket for å øke forskjellen i tidspunkt for stengelstrekning mellom de to gruppene. Seks uker (42 dager) etter snittet for strekningstidspunkt i hver av de to gruppene (hver 52 stk) ble det telt antall skudd i strekning, antall skudd totalt og om de hadde synlig knopp. Målingen var primært tiltenkt bare de 52 tidligste og seneste men inkluderte alle de 96 seneste og tidligste for å samle mer data. Tidspunktet for denne registreringen var 20.11.16 for den tidlige gruppen og 1.1.16 (målt 2.1.16) for den sene gruppen. Andelen skudd i strekning ble beregnet (Skudd i strekning/totalt antall skudd). Etter registreringene 6 uker etter snittet for DTS i den tidlige gruppen ble de respektive plantene klippet ned til 3-5 cm for å bekjempe meldugg, gjøre plantene enklere å håndtere og så de tok mindre plass frem til høsting av blader for DNA ekstraksjon. Den tidligste gruppen ble flyttet til et annet vekstrom i perioden 10.12.15 til 4.1.16 for å gjøre vegetasjonen bordene mindre tett.

Plantene som ennå ikke hadde startet stengelstrekning 4.12.15 (146stk) ble klippet ned 10.12.15 og flyttet sammen på et bordene i drivhuset under samme forhold som tidligere i forsøket. 12 planter som ikke var strukket etter registreringene var avsluttet (4.12.15) ble observert å ha startet stengelstrekningen 19.1.16. Disse plantene tas ikke med til GBS men de ble overvintret ved kald temperatur (frostfritt) i drivhuset. Påfølgende sommer blomstret samtlige planter.

Avvik

2 planter ble kastet med uhell (begge 94 DTS)**.** Noen planter som egentlig hørte til den sene gruppen ble ikke med blant de 52 utvalgte hvorav 2 individer var sterkt forkrøplet og hadde en type spraglet klorose (74DTS og 71DTS), påvirkningen på veksten har trolig innvirkning på blomstringsmønsteret så det ble valgt vekk, Totalt 5 planter viste en abnormalitet av spraglet klorose. Utover dette viste 2 andre abnormal vekst. En av plantene med spraglet klorose ble med i GBS og tilhører den sene gruppen.

Enkelte planter hadde et trekk som var svært treg strekningshastighet. Disse dannet tydelige strekningsstengel (1-2 cm) i normal tid med elongerte ikke før etter flere uker, og selv etter strekningen begynte gikk det svært langsomt. Plante nummer 151 og 168 er to ekstremer som ble valgt bort på

grunn av usikkerhet om strekningstidspunkt. Det ble først oppdaget meldugg 30.10.15. Det var synlig meldugg på 27,8% av plantene 1.11.15. Dette spredte seg til hele rommet. Plantene ble sprøytet 10.12.15 med Confidor. 10.11.2015 ble det observert bladlus. Snylteveps ble sluppet ut 23.11.2015. To ulike arter, *Aphidus colemani* og *Aphidius ervi* ble brukt for å være sikrere på at de var kompatible med typen bladlus. 30.10.15 ble det observert apotecier på jorda i 42 av pottene. Det var bare observert mellom plante 231 og 383, et spenn på 152 planter bare på ett og et halvt bord. Antageligvis stammer soppen fra en av sekkene med jord. Det stemmer godt med at et 80 L sekk holder til omtrent 145 potter ved 550 cm^3 per potte. Plantene så ikke ut til å være påvirket av soppen.

2.3 Statistisk analyse a fenotypiske data

Analyser ble gjort med statistikkprogrammet Minitab. Standardavviket i dager til strekning (DTS) innen hver gruppe, tidlige og sene ble regnet ut. Tidlige og sene ble sammenlignet med toveis T-test hvor null hypotesen var likhet i DTS for å bekrefte at de var ulike og på gruppenes av andel skudd i strekning som ble målt 42 dager etter gruppens gjennomsnittlige DTS. Enveis anova på bord ble utført for å undersøke om plassering hadde påvirkning på DTS, kanskje forårsaket av variasjon i naturlig lysinnstråling, romlige temperaturforskjeller, vanningsforskjeller eller annet.

Ved statistisk analyse av om plassering hadde noe å si på tidlighet så ble alle plantene som ikke strakk seg innen 4.12.15 og 5 forkrøplete planter tatt ut av beregningen. Det ville være vanskelig å bestemme en tallverdi som er praktisk å jobbe med. De er heller ikke hensiktsmessig å ha med fordi deres krav for strekning aldri ble møtt, da de ligger langt utenfor normalfordelingen av de som gikk i strekning under forholdene i drivhuset.

2.4 Sammendrag av prosessen i lab

Isolering av DNA

Bladmateriale ble i første omgang høstet til 2 stk 96 plater og ekstrahert med DNeasy 96 plantkit (QIAGEN). DNA kvalitet ble testet på agarosegel neste dag. DNA konsentrasjon ble kvantifisert med Victor 3 Multilabel 1420 (Picogreen™ metode) med pipetteringshjelp fra en Beckman coutier robot. 24 prøver ble vurdert til å inneholde for liten mengde DNA og de respektive individene ble isolert på nytt med Qiagen DNeasy minikit og kvantifisert med Qubit. Ni tilfeldige prøver fra første eluering med 96kit ble også kvantifisert med og Qubit for få et bilde av hvordan disse relaterer seg til hverandre. Forholdet

mellom Picogreen/QuBit estimert til 1,37. På bakgrunn av dette ble konsentrasjonene (QuBit estimat) i de 24 prøvene ekstrahert med «plant minikit» ganget med 1,37. Kuttbarhet med HindIII ble kontrollert med inkubering i 2 timer på 37°C med 1U enzym / 100ng DNA i totalt 15µl volum (vedlegg 2 tabell 1). på stikkprøver av isolering med DNeasy 96 plantkit og DNeasy minikit.

2.5 Pooling

De 52 plantene i hver gruppe (Tidlige og Sene) ble tilfeldig fordelt i tre undergrupper. DNA i hver undergruppe ble kombinert så det ble 3 replikate «DNA-pooler» fra de tidlige plantene og 3 fra de sene plantene. Konsentrasjon ble kompensert med volum slik at hvert individ bidro med lik mengde DNA i poolen (1400 ng). Det var det høyeste mengde mulig på grunn av begrenset tilgjengelig mengde isolert DNA fra enkelte individer. På grunn av pipetteringsfeil ble det overført ca 10% for mye DNA fra to individer i «sene 1». Det kommer tydelig frem på standardavviket for individuelt bidrag målt i DNA i pool (Tabell 1). Konsentrasjonen av DNA i poolene lå mellom 19,4 - 25,3 ng/µl. Anslått volum i pool ble delt på antall brønner over de to platene den skulle fordeles på +1,5 eller 2 slik at det skulle rekke til alle brønnene, med en liten margin i tilfelle pipetteringsfeil. Det totale volumet på polene beskrevet i tabell 1 er basert på hva som ble pipettert(Tabell1).

Tabell 1 Oversikt over fordeling av DNA i de seks Poolene, S tidlige og 3 Sene. Fordelt over 2 plater X 95 brønner(190 totalt). Summen av mengden DNA i poolen er basert på den pippeterte mengden DNA isolat korrelert med estimat av konsentrasjonen basert på kvantifisering med Picogreen eller Qubit. Stdev av individbidraget er basert disse tallene som er rundt 1400ng, med to outliers i «Sene 1» på grunn av pipetteringsfeil.

DNA Pool	Tidlige 1	Tidlige 2	Tidlige 3	Sene 1	Sene 2	Sene 3
Antall individer	17	17	18	17	17	18
Volum av pool (µl)	1242	1168	1050	1074	943	1066
Stdev individbidrag ng DNA til pool	6,6	4,3	4,9	45,2	3,6	4,5
Antall brønner, plate (Apek1/Pst1)	15/16	16	16	16/15	16	16
Mengde DNA per brønn (ng)	737	704	746	732	708	746
konsentrasjon (ng/µl)	19,4	20,4	24,08	20,99	25,27	23,69
Volum per brønn til plate (µl)	38	34,5	31	33	28	31,5

DNA fra alle 6 pooler ble fordelt i to plater med 96 brønner (15 eller 16 brønner per pool +1 blank) (vedlegg 3, tabell 1 og 2). Prøvene ble sendt til på tørris til Biotechnology Resource Center, Cornell University, for GBS.

2.6 Behandling hos Biotechnology Resource Center, Cornell University

Preparering av GBS-biblioteker, Sekvensering ble utført av Genomic Diversity Facility, Cornell University. En plate ble kuttet med Pst1 og den andre ble kuttet med Apek1 før sekvensering. Her oppsummerer jeg kort stegene i Tassel-GBS 'Discovery Pipeline'(Glaubitz et al. 2013) som ble brukt til å behandle rådataene. Denne prosessen er visualisert i et flytskjema (Vedlegg 6, Figur 1). Lengden på råsekvensene i FASTQ-filen ble trimmet til 64bp (det inkluderer ikke barcoden). Tag-sekvensene i FASTQ-filen(e)(det ble to for Apek1) ble alignet med hverandre for å telle hvor mange ganger like sekvenser var lest (Master Tagcounts). Alle under tre observasjoner ble kastet på dette nivået. «Master tagCounts» ble alignet til rødkløver-genomet (DeVega et al. 2015) (redclover_v2.1) og informasjonen ble skrevet til filen «TagsOnPhysicalMap(TOPM)». «Master TagCounts» ble igjen brukt med FASTQ filen(e) for å se distribusjonen av master tags i samplene ved å bruke barcode-nøkkelfilen til å linke prøveID til tagsekvens. Informasjonen ble skrevet til filen «TagsByTaxa». TagsOnPhysicalMap og TagsByTaxa ble brukt sammen til å finne SNPer. SNPene fikk også lagt til informasjon om allel og i TagsOnPhysicalMap filen. I tillegg ble det i en separat .vcf gjort grovfiltrering på å fjerne minor allele frequency <0,01, missing data per site >90% (proporsjon av sampler med felles tags). Filtrert og ufiltrert data ble levert tilbake sammen med en rapport med detaljer om GBS-pipeline prosessen (utdrag i vedlegg 6). Tabeller som viser dybde og missingness ble laget for både den rå og den grovfiltrerte .vcf filen (Vedlegg 6, Tabell 4 og 8). Det ble også konstruert grafer over distribusjon av minor allelfrekvens i bialleliske loci (Vedlegg 6, Figur 2 og 4) og multi dimensional scaling plot (MDS) over bialleliske SNPer (Vedlegg 6, Figur 3 og 5). Begge laget med VCFtools versjon v0.1.12a og PLINK versjon v1.07 og tar utgangspunkt i den grovfiltrerte .vcf filen.

2.7 Databehandling

2.7.1 Filtrering med Pearlscript

Den grovfiltrerte .vcf filen All.filtered.recode.vcf.gz ble pakket ut med «gunzip» og filnavnet ble endret fra .vcf til .txt. Et filter kodet i perl (se vedlegg 5) gjør følgende: Lokaliserer verdiene for antall reads major og minor allele. Disse kalles i .vcf filten «Allelic depth», separerer med TAB og fjerner resten av verdiene rundt. Under vises et eksempel hva hvordan data for en biallel og triallel SNP ser ut og hvilke som hentes herfra.

Biallelisk:	0/1:92,79:171:100:255,0,255	→	92	79	
Triallelisk:	0/2:3,1,2:6:99:57,0,93	→	3	1	2

Skriptet definerer hvilke av de 95 rørene som tilhører hvilken pool, deretter nulles antall reads for major og minor allele ut i SNP/rør kombinasjon der hvor minst ett av allelene har reads på 127. Dette gjøres for at metning ikke skal påvirke beregning av allelfrekvensen. Maksimalt antall reads per tag per taxon (reads per tag per rør) er 127 i Tassel 3 «discovery» pipeline (Jeff Glaubitz 2013). Videre summerer skriptet separat for hvert allel over alle 95 og denne SNP filtreres bort dersom MAF (Minor allele frekvency) er <0,05 (For trialleliske SNP ble MAF regnet som den samlede frekvensen av begge de to «minor» allelene). Siste steg i filteret fjerner SNPer hvor summen av antall reads for alle (to eller tre) alleler en pool er mindre enn 100.

2.7.2 Utregning av allelfrekvens og Fst

Den nye .txt tekstfilen generert av pearlskriptet ble importert i MS Excel som ble brukt til databehandling videre. Allelfrekvensen for alle SNP ble regnet ut først for hver av de 6 poolene ved å summere antall reads til de to allelene. Allelfrekvensen i den tidlige og sene gruppa ble og regnet ut baser på summerte antall reads av alle tre pooler i hver gruppe. Det ble filtrert med kriteriet at forskjell i allelfrekvens fra hver gruppe måtte gå i samme retning i forhold til snittet av den andre gruppen. Grad av genetisk differensiering mellom subpopulasjoner kan beskrives som fikserings indeksen Fst, først definert av (Wright 1951). Fst kan variere fra 0 (identisk) til 1 (helt ulik). Ut fra allelfrekvensene til poolene ble det regnet Fst for alle SNP etter formelen på s.386 i Genetics of populations (Hedrick 2011).

$$Fst = \frac{\overline{q^2} - \bar{q}^2}{\bar{q}(1-\bar{q})} = \frac{V(q)}{\bar{q}(1-\bar{q})}$$

For hver enkelt SNP ble seks Fst verdier først regnet ut. Disse burde være like innen hver DNA pool dersom resultatene er reproduserbare og pålitelige. Deretter ble hver enkelt pool den ene gruppen testet mot hele (snittet av) den andre gruppen og omvendt. Videre ble 2N Fst regnet ut ved å multiplisere Fst verdien med 2 ganger antall individer i Fst testen. 2 x (Npool + Ngruppe) x Fst = x^2 Tallene ble chi-square testet mot et signifikansnivå på P<0,1 , P<0,05 og P<0,01. SNP skal ha signifikans i alle seks retninger for at den skal regnes som signifikant. Dette reduserer antallet falske positive betraktelig ved at fordi det er usannsynlig at Fst blir signifikant i alle seks parvise sammenligninger. For triallele SNP ble Fst utregningen i tillegg gjort i tre omganger der q var summen av en eller to alternative

baser som ble testet mot den tredje.

FDR (False discovery rate) ble regnet ut som:

$$fdr = \frac{Antall\ FST\ testet * p^3}{Antall\ positive}$$

FDR er sjansen for at SNP på det gitte nivået er en falsk positiv. Det ble satt et kriterie at FDR skulle være lavere enn 5 %. Derfor tas ikke noen SNP med signifikansnivå på kun 0,1 med videre fordi FDR var over 0,05 (Tabell 4).

2.8 Videre behandling

Filene med SNP informasjon all.filtered.vcf ble sortert med tassel med «SortGenotypeFilePlugin». Det var første steg som var nødvendig for å få vise filen i Tassel. «GenotypeSummararyPlugin» ble kjørt. Forøvrig ble Tassel ofte brukt til å visuelt utforske SNPer. SNP med signifikant forskjellig allelfrekvens i de to fenotypiske gruppene tidlige og sene ble plottet inn på kromosomer i MapChart. SNP med P<0,01 ble letet opp i genombrowserene Jbrowse og Gbrowse på nettsiden Legume information system. Disse viste nukleotidsekvensen og predikerte genmodell basert på referansegenomet. Genombrowserene viste også om området var syntenisk med områder i gruvesneglebelg (Medicago truncatula) eller kikert (Cicer arietinum).

http://legumeinfo.org/genomes/jbrowse/?data=TP2.1

http://legumeinfo.org/genomes/gbrowse/Tp2.1

Gjennomsnittlig LD over 100 kb er 0,15-0,25 rødkløver (DeVega et al. 2015). Det er en god pekepinn på hvilken avstand kandidatgener kan befinne seg innenfor. De tre SNPene med størst differanse i allelfrekvens mellom tidlig og sen gruppe ble undersøkt grundigere ved å notere avstand og id til nabogener i området rundt. Størrelsen på området ble satt til å være et vindu på 25 kpb (12,5 kb i hver retning). Det undersøkt om noen SNPer med signifikans (P<0,05) så ut til å ligge i samme område på kromosomet som QTLer som ble kartlagt for blomstringstid i rødkløver i studien av Herrmann et al (2006), QTL i *M. truncatula* knyttet til blomstringstid (Pierre et al. 2008) og QTL for vinteroverlevelse i Rødkløver beskrevet av Klimenko et al. (2015) fordi vinteroverlevelse kan ha tilknytning til tidlighet.

Ved hjelp av et pearl script ble det hentet ut de faktiske sekvensene som korresponderte til signifikante SNP fra 2088.sam filen. Det måtte hentes ut på denne måten fordi filen var for stor til å åpne hele i excel eller vanlig tekstbehandler. Posisjonen til en SNP var ikke det samme som navnet til fragmentet derfor

ble det søkt med nummeret til SNP +/- 64 bp, som er lengden på en tagsekvens, i en linkage group av gangen. Ulempen med denne metoden var at det var umulig å vite hvilke tagsekvensene fra samme område som var mest representert. Ofte kan det være så mye som 50 ulike tagsekvenser fra samme område. Det viste seg forøvrig raskt at det kun var nødvendig å søke tilbake til maksimum – 64 bp ettersom tagsekvensene allerede var alignet til referansegenomet og nummerert i stigende rekkefølge så det spilte ikke lenger noen rolle hvilken vei de var sekvensert.

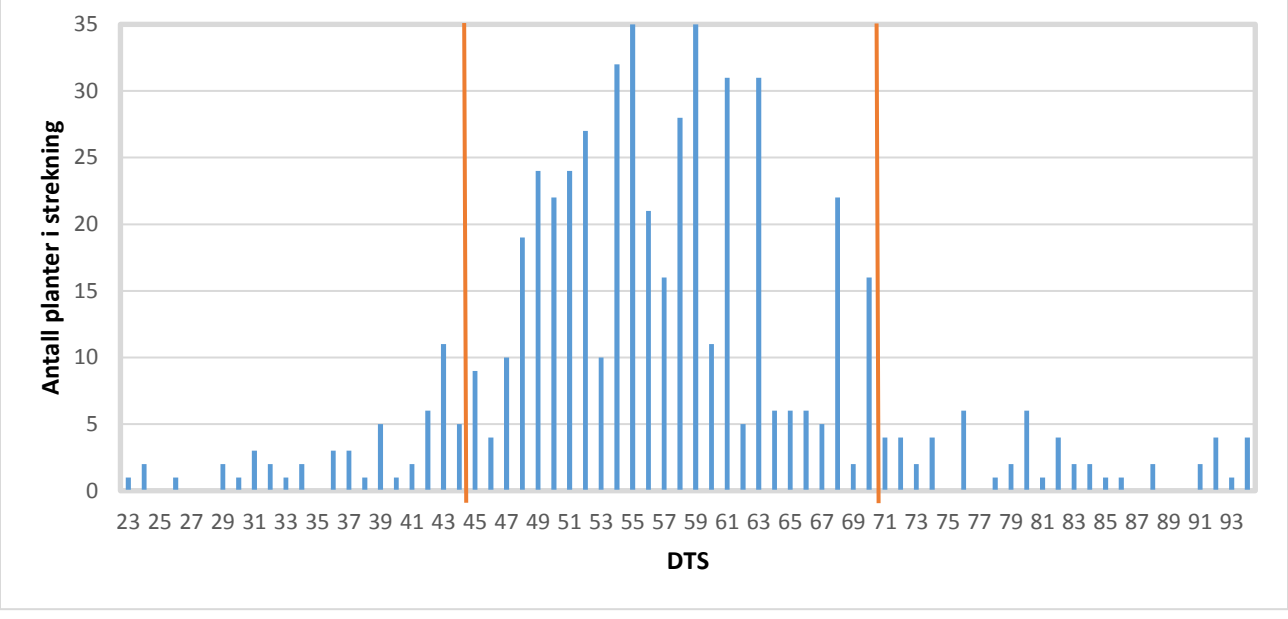
Dersom sekvensen ikke lå i noe gen ble det utført blast-søk via LIS og NCBI. Ved blasting var det mulig å vise regionen i Gbrowse. Visningsvinduet ble utvidet og deretter lastet ned nukleotidsekvens i visningsområdet som kunne åpnes i MEGA6.

Blastsøk ble utført via ncbi https://blast.ncbi.nlm.nih.gov/
og via LIS Legume information system http://legumeinfo.org/blast/nucleotide/nucleotide
Alignments av tagsekvenser ble gjort med EMBOSS Water på http://www.ebi.ac.uk/

3. Resultater

3.1 Fenotyping

DTS (dager til strekning) var tilnærmet normalfordelt og er basert på data er korrigert for sen registrering (Figur 1). Toppene på figuren skyldes at det går noen dager mellom hver registrering. Den første gruppen strekker seg frem til 44 DTS og den sene gruppen strekker seg fra 71 DTS. Første tidspunkt for registrering var 22-23 oktober. Alle individene i den «tidlige gruppen» n=52 ble registrert da. Av disse hadde over halvparten allerede mer enn 15 cm lang stenger og ingen hadde stengeler som var strekt mindre enn 6 cm (Vedlegg 1, tabell 1).



Figur 1: Frekvensfordeling av antallet dager til stengelstreking (DTS med korrigering for sen registrering). Plantene ble dyrket ved 20 timer daglengde og 16°C. 146 av totalt 672 planter hadde ikke begynt å strekke seg da registreringene ble avsluttet etter 94 dager. De 52 tidligste og 52 seneste plantene (indikert før og etter oransje streker) ble tatt videre til genotypisk karakterisering, samt fenotypisk karakterisering av skudd 6 uker etter gruppenes gjennomsnittlig tidspunkt for stengelstrekning.

Blant alle plantene som startet strekningsfasen innen 4.12.15 (94 DTS) var gjennomsnittlig DTS 57,25 dager. Den tidlige og sene gruppen hadde signifikant forskjellig DTS (t-verdi=32,81 P<0,0001) (vedlegg 1: tabell 5). Den tidlige gruppen hadde gjennomsnittlig DTS på 37,73 dager med standardavvik på 6,03. Den sene gruppen hadde gjennomsnittlig DTS på 80,52 med standardavvik på 7,33. Forskjell i DTS mellom de 52 tidlige og de 52 sene plantene illustreres godt i et boxplot (vedlegg 1: Figur1). Enveis Anova for å teste om bordplassering hadde innvirkning på DTS ga P=0,347, altså ingen innvirkning (vedlegg 1, Tabell 7). Seks uker etter det gjennomsnittlige strekningstidspunktet for hver av gruppene ble antall vegetative

og strekte skudd talt opp (vedlegg 1 Tabell 1, Figur 4, figur 5). Seks uker vil si 20.11.15 i den tidlige gruppen og 2.1.16 i den sene gruppen. Ved disse tidspunktene hadde 48/52 av de tidlige plantene, og 10/52 av de sene blomst/synlig knopp. De tidlige plantene hadde signifikant høyere snitt av andel skudd i strekning. Andelen var nesten dobbelt så stor (vedlegg 1, figur 5). T-test mellom den tidlige og sene gruppen i andel skudd i strekning (skudd i strekning/totalt antall skudd) oppga signifikant forskjell med P<0,0001 (vedlegg 3 tabell 6). I de 52+52 som skulle karakteriseres genetisk var snitt av andel skudd i strekning var 0,7 (stdev 0,16) blant de tidlige og 0,39 blant de sene (stdev 0,13) (vedlegg 3, figur 5). Til sammenligning var snittet blant de 95 tidligste 0,67 (stdev 0,15) og i de 95 seneste 0,43 (stdev 0,16). Med unntak av noen outliere var det en sammenheng mellom strekningstidspunkt og andel skudd i strekning (Vedlegg 1, figur 5). Det ble og sammenlignet antallet vegetative skudd som var 2,88 ganger høyere i den sene i forhold til den tidlige gruppen (vedlegg 1 figur 4). Andelen planter med blomst målt seks uker gjennomsnittlig DTS var mindre desto lengre tid plantene brukte på å starte stengelstrekning (Vedlegg 1, Figur.

3.2 Resultater av GBS fra Cornell University

Informasjonen i dette kapittelet er hentet fra GBS-rapportene fra Cornell University (vedlegg 6 og 7)

Apek1

Robotfeil i en av radene et sted i prosedyren førte til at 25 av samplene ble kjørt i en egen «lane» og sluppet i separat fastq og barcode-nøkkel. Etter merging med minimum 3 reads per tag ble antall tags 2788388 i Apek1. Alignment av tags fra Apek1 med rødkløver-genomet resulterte i 284967 SNPer. Filtreringen Cornell utførte (spare på bare de med minor allele frequency >0,01, missing data per site <90%) resulterte i 276576 SNPer. Gjennomsnittlig dybde for hver prøve etter filtreringen var 49,974 med standardavvik på 9,27 (vedlegg 6). Det vil si en gjennomsnittlig dybde i hver SNP på 4748.

Pst1

Etter merging med minimum 3 reads per tag ble antall tags 1237695 i pst1. Alignment av tagssekvenser fra Pst1 mot referansegenomet resulterte i 34713 SNPer. Filtreringen Cornell utførte (minor allele frequency >0,01, missing data per site <90%) reduserte dette til 32553 SNPer. Gjennomsnittlig dybde for hver prøve etter filtreringen var 17,245 med standardavvik på 4,15(Vedlegg 7). Det vil si gjennomsnittlig dybde i hver SNP på 1638.

Diagrammer i vedlegg

Frekvensfordelingen av MAF i bialleliske loci viser at majoriteten av SNPer har MAF under 0,05 i begge

enzymene (vedlegg 6, Figur). Mengden MAFer større i Apek1 enn i pst1. Frekvensen av MAF var lavest mellom 0,3-0,45. I to multi dimensional scaling (MDS) plot (se vedlegg 6 og 7 side 10) adskilles de tidlige og sene gruppene i begge enzymer vertikalt i pst1 og horisontalt i apek1.

3.3 Databehandling av GBS data

De filtrerte dataene fra Cornel «all.filtered.vcf» ble filtrert videre med et pearlscript som resulterte i at det på data fra Apek1 reduserte antall SNPer til 43242 SNP, det er en reduksjon etter alignment på 84,98%. Filtreringen med pearlskriptet på data fra Pst1 reduserte antall SNPer til 3590. Reduksjon fra alignment til etter pearl er på 89,78%. Gjennomsnittlig ble det med Pst1 funnet 1 SNP per 117 kb mens det med Apek1 ble funnet gjennomsnittlig 1 SNP/9,7 kb (Tabell 2). Tettheten av SNP etter filtreringen er over ti ganger høyere med Apek1 enn med Pst1. Fordelingen av SNPer etter filtrering varierte mellom kromosomene. Det var mindre variasjon i tetthet SNPer funnet med Apek1 enn med Pst1 (Tabell 3).

Tabell 2 Reduksjon av antall SNPer etter filtreringer. Tettheten av SNPer etter pearlfiltrering er baset på genomstørrelse på 420 Mb

GBS enzym	Antall SNP etter alignment	Cornell filter	Etter Cornell filtrering	Pearl filter	Etter pearlfilter	Gjennomsnittlig tetthet etter pearlfilter
Pst1	34713	*-2160*	32553	*-28963*	3590	1 SNP/ 117 kb
Apek1	284967	*-8391*	276576	*-233334*	43242	1 SNP / 9,7 kb

Tabell 3 Antall SNPer plassert på kromosomer og tettheten i forhold til hele kromosomstørrelsen. Dette er alle SNPer som var igjen eller filtrering med pearlfilteret (MAF>0,05 og, fjernet prøver der reads for en av allelene nådde 127, og krav på minst 100 reads i hvert av seks pooler)

Kromosom		Tp1 28,14 Mb	Tp2 32,56 Mb	Tp3 31,06 Mb	Tp4 28,91 Mb	Tp5 15,27 Mb	Tp6 22,68 Mb	Tp7 30,55 Mb
Pst1	SNPer	376	413	430	278	190	263	330
	SNP tetthet	1/74,84 Kb	1/78,84 Kb	1/72,24 Kb	1/103,98 Kb	1/80,38 Kb	1/86,25 Kb	1/92,58 Kb
Apek1	SNPer	4333	4844	4637	3890	1953	3522	4129
	SNP tetthet	1/6,49 Kb	1/6,72 Kb	1/6,70 Kb	1/7,43 Kb	1/7,82 Kb	1/6,44 Kb	1/7,40 Kb

I excel filene fra generert fra filtrert data fra pearlfilteret ble først biallele og triallele SNPer delt i to ulike regneark. Andelen triallele SNP etter filtreringen var omtrent lik med begge enzymer, Pst1 (0,0652) og

Apek1 (0,0660). Det ble regnet Fst verdier for de 3590 + 43242 SNPene. Formler i Excel testet om 2N av Fst verdiene i alle 6 pooler testet mot motsatt gruppe hadde chi-square signifikans på <0,1 nivå, <0,05 nivå eller <0,01 nivå. (Tabell 4). Det ble regnet fdr på antall funnet på ulike signifikansnivåer. SNPer funnet med Pst1 hadde lavere *fdr* på alle signifikansnivå sammenlignet med apek1. Fdr var høy, over 0,05 på SNPer med signifikans på kun P<0,1 og derfor ble de forkastet. På P<0,05 ble på biallele SNP fra Pst1 redusert til 37 stk og fra Apek1 160 mens på P<0,01 nivå var det redusert til 15 stk fra pst1 og 42 stk fra apek1. Proporsjonalt fant Pst1 flere signifikante biallele SNPer enn Apek1 på P<0,01 nivå i forholdet til antallet SNP etter pearl-filteret (Tabell 4) (Pst1: 0,0045) og (Apek1: 0,0010). Henholdsvis 40% (Pst1) og 38% (Apek1) av SNPene med signifikante forskjellig allelfrekvens i tidlige og sene planter på P<0,1 nivå ligger i scaffolds. Litt over halvparten (24/45) av triallele SNP på P<0,1 funnet med Apek1 ligger i kromosomer mens fra pst1 ligger alle med P<0,1 (4 stk) i kromosomer. Blant triallele ble det på signifikans på P<0,05 nivå funnet 3 SNPer med pst1 og 19 SNPer med apek1. Dybden av reads per SNP varierte mye og det så ikke ut til å ha noe sammenheng med signifikansnivået (Tabell 4, 5, 6).

Enkelte av SNPene som ble funnet representerte samme posisjon, for eksempel lengre indels. De 37 signifikante SNPene funnet med Pst1 på P<0,5 nivå kan reduseres til 29 dersom suksessive SNP av indels tas ut og gjøres det samme med de på P<0,01 nivå er det igjen 12 SNPer med signifikant forskjellig allelfrekvens (vedlegg 4, tabell 2, 3, 4). Noen av de trialleliske SNPene var signifikant i to retninger (Vedlegg 3, Tabell 4).

Tabell 4 Antall Signifikante biallele SNP etter «pearskript» (MAF>0,05 og, fjernet prøver der reads for en av allelene nådde 127, og krav på minst 100 reads i hvert av seks pooler). Tabellen grupperer SNPer etter signifikansnivået av Fst verdiene sammenligning av allelfrekvensene i DNA pooler fra planter med tidlig eller sent tidspunkt for strekning, samt signifikansnivåets gjennomsnittlige antall reads(dybde). (Alle innunder signifikansnivået er fremstilt sammen, slik at P<0,01 omfattes av P<0,05 som igjen omfattes av P<0,1)

GBS enzym	Biallele SNP	Signifikante på P<0,1 nivå	Signifikante på P<0,05 nivå	Signifikante på P<0,01 nivå
Pst1	3356	55 (fdr: 0,061)	37 (fdr: 0,0113)	15 (fdr: 0,00022)
Snitt dybde	4187	4842	3483	4663
Apek1	40387	321 (fdr 0,126)	160 (fdr: 0,0315)	42 (fdr: 0,00096)
Snitt dybde	3088	2479	2354	1914

Tabell 5 Antall Signifikante biallele SNP etter «pearskript» (MAF>0,05 og, fjernet prøver der reads for en av allelene nådde 127, og krav på minst 100 reads i hvert av seks pooler). Tabellen grupperer SNPer etter signifikansnivået av Fst verdiene sammenligning av allelfrekvensene i DNA pooler fra planter med tidlig eller sent tidspunkt for strekning, samt signifikansnivåets gjennomsnittlige antall reads(dybde). (Alle innunder signifikansnivået er fremstilt så P<0,01 omfasses av P<0,05 som igjen omfattes av P<0,1)

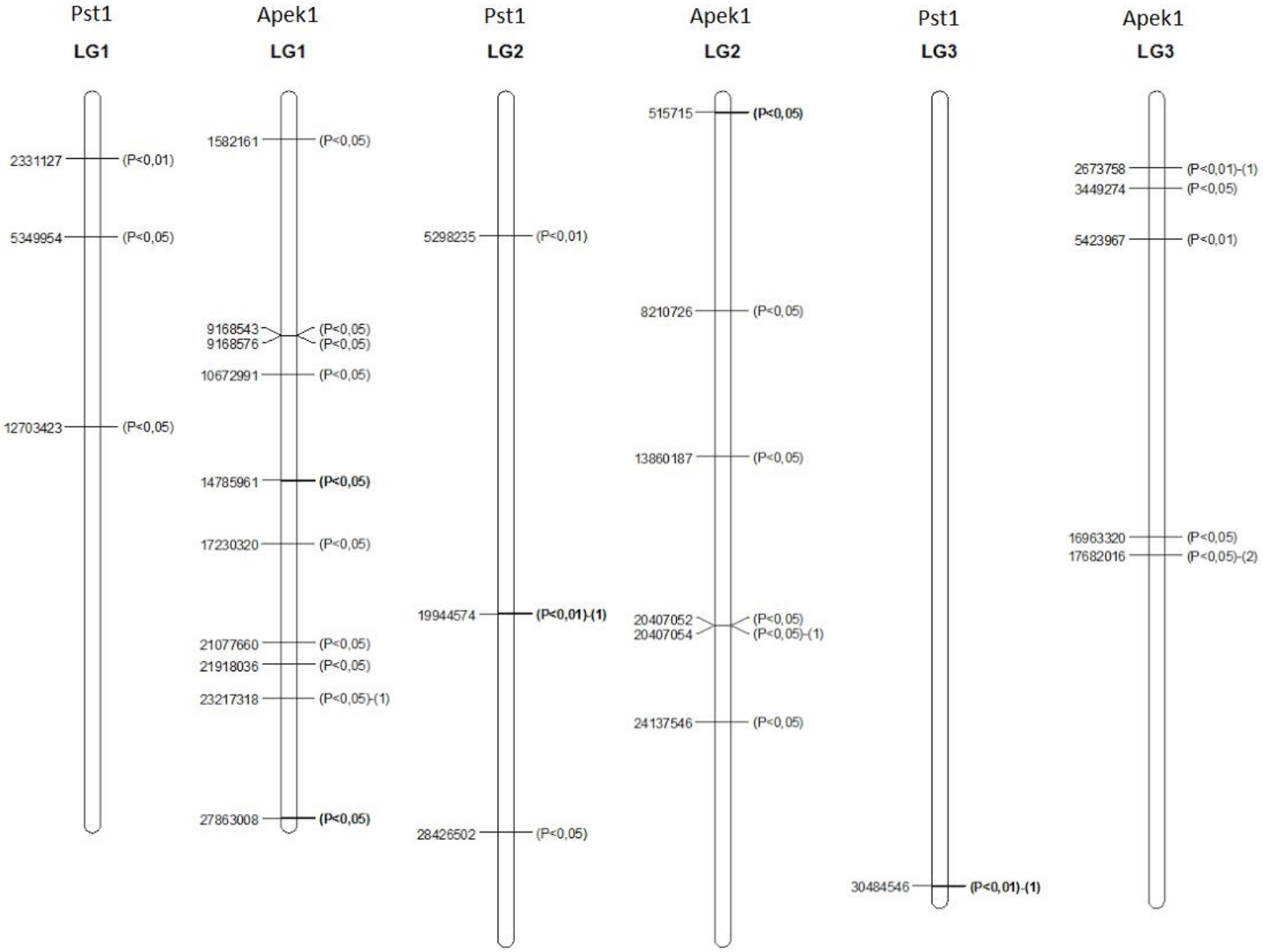
GBS enzym	Triallele SNP	Signifikante på <0,1 nivå	Signifikante på <0,05 nivå	Signifikante på <0,01 nivå
Pst1	234	4 (fdr:0,0585)	3 (fdr: 0,00975)	2(fdr: 0,000078)
Snitt dybde	*793*	*6310*	*6310*	*6900*
Apek1	2855	45 (fdr: 0,063)	19 (fdr: 0,01878)	4 (fdr: 0,000713)
Snitt dybde	*2780*	*2257*	*1972*	*1654*

Dybden er påvirket av at enkelte prøver av SNP er fjernet ved filtrering *maks 127 reads*

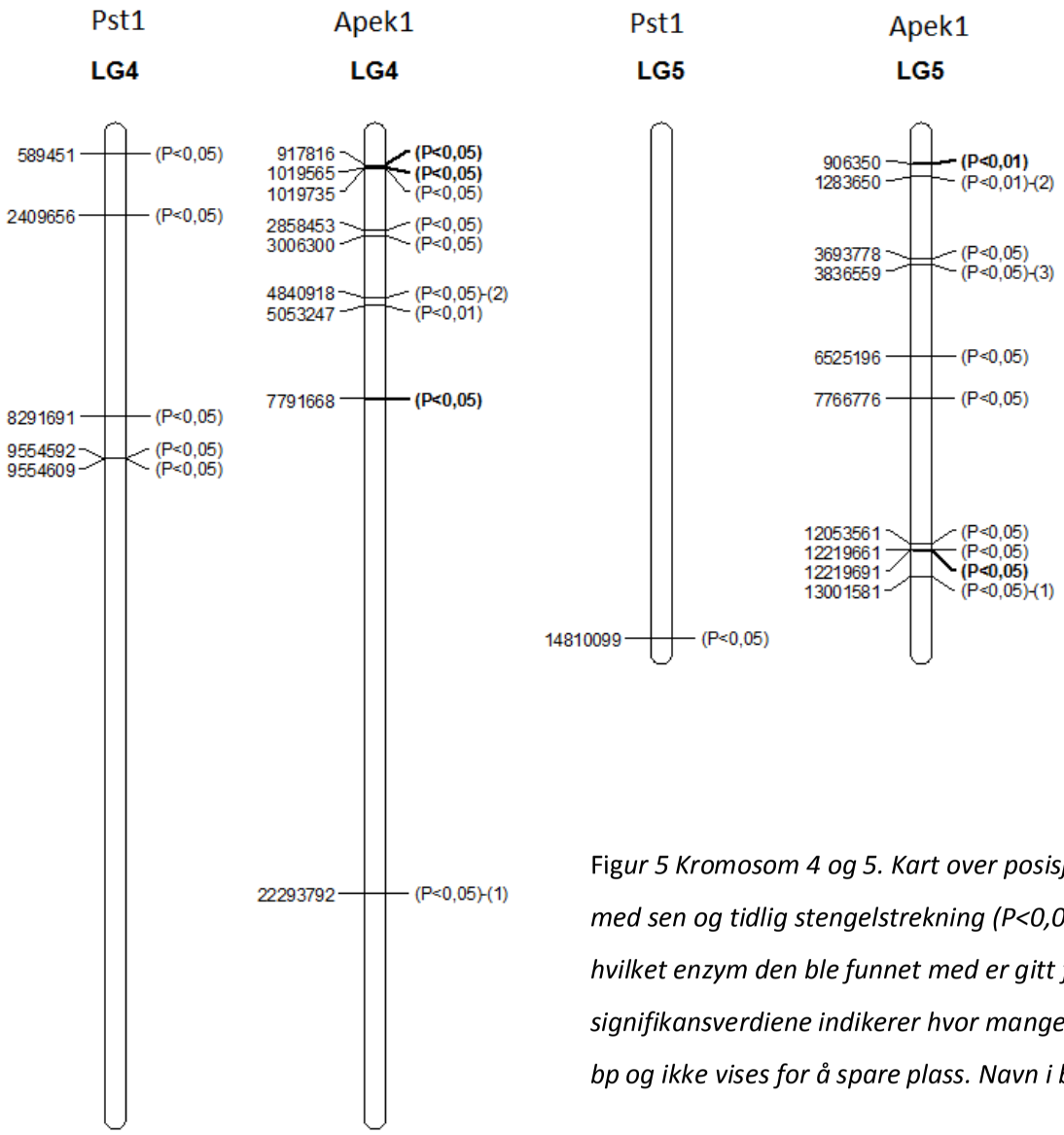
3.4 Videre behandling av SNPer med signifikant ulik allelfrekvens i de to fenotypiske gruppene (P<0,05)

3.4.1 Oversikt over signifikante SNPer

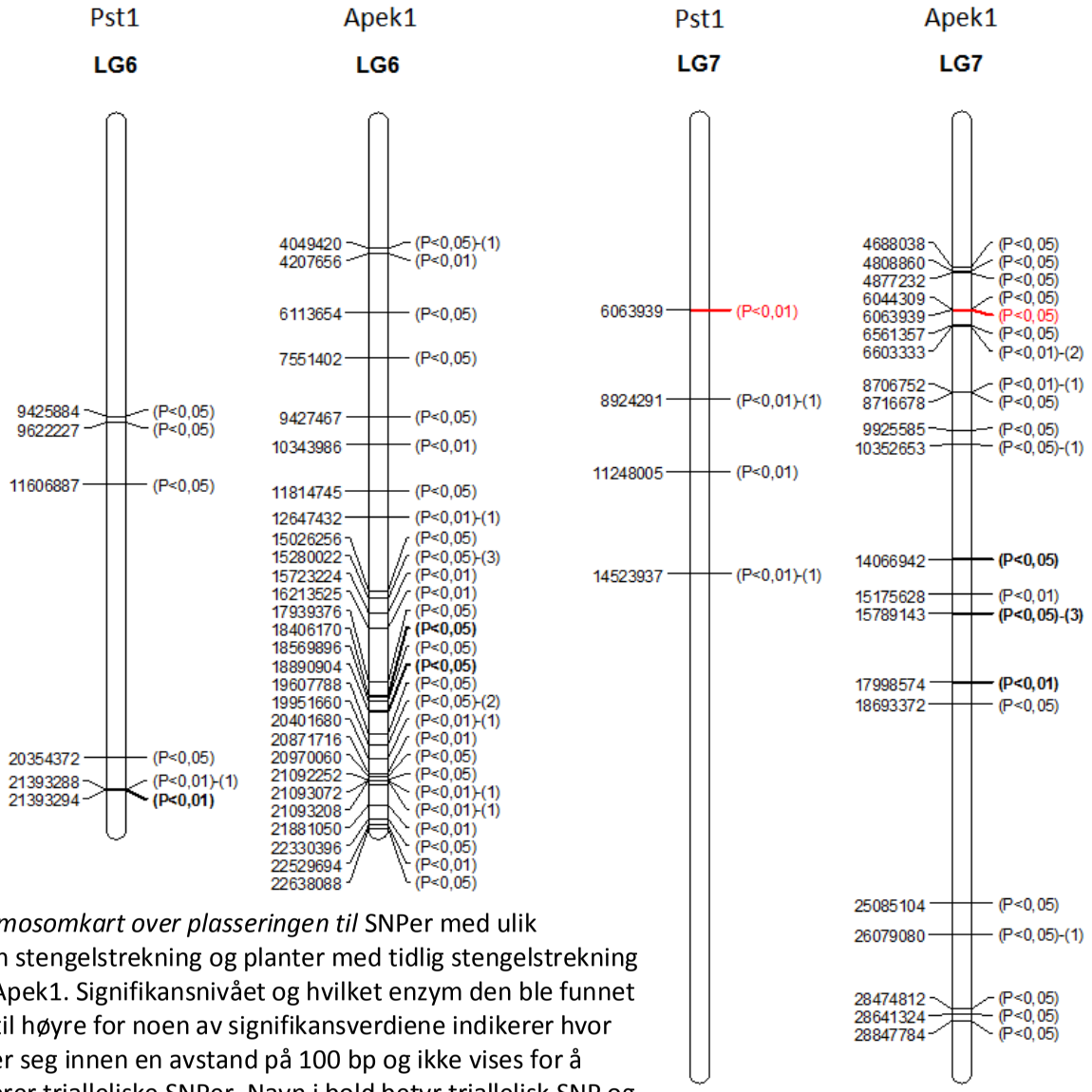
Kart over kromosomposisjonene til SNPer med ulik allelfreksens mellom gruppene med tidlig og sent tidspunkt for stengelstrekning med signifikans på minst (P<0,05). Kartene viser distribusjonen over de syv kromosomene fremvist separat for de to GBS omgangene med enzymene pst1 tv. og apek1 th. (Figur 4, 5, 6). Det er en spesielt høy konsentrasjon av SNPer funnet med apek1 i nedre del av kromosom 6 (Figur 6). Mesteparten av de signifikante SNPene i kromosomer på P<0,01 nivå lå i gener (Vedlegg 4 Tabell 1).



Figur 4 Kromosom 1, 2 og 3. Kart over posisjonen til SNPer med ulik allelfrekvens mellom planter med sen og tidlig stengelstrekning (P<0,05) funnet med Pst1 og Apek1. (P<0,05) *Signifikansnivået og hvilket enzym den ble funnet med er gitt for hver SNP. (#), til høyre for noen av signifikansverdiene indikerer hvor mange nabo-SNP som befinner seg innen en avstand på 100 bp og ikke vises for å spare plass. Navn i bold indikerer trialleliske SNPer.*



Figur *5 Kromosom 4 og 5. Kart over posisjon til SNPer med ulik allelfrekvens mellom planter med sen og tidlig stengelstrekning (P<0,05) funnet med Pst1 og Apek1. Signifikansnivået og hvilket enzym den ble funnet med er gitt for hver SNP. (#), til høyre for noen av signifikansverdiene indikerer hvor mange nabo-SNP som befinner seg innen en avstand på 100 bp og ikke vises for å spare plass. Navn i bold indikerer trialleliske SNPer.*



Figur 6 Kromosom 6 og 7. Kromosomkart over plasseringen til SNPer med ulik allelfrekvens i planter med sen stengelstrekning og planter med tidlig stengelstrekning (P<0,05) funnet med Pst1 og Apek1. Signifikansnivået og hvilket enzym den ble funnet med er gitt for hver SNP. (#), til høyre for noen av signifikansverdiene indikerer hvor mange nabo-SNP som befinner seg innen en avstand på 100 bp og ikke vises for å spare plass. Navn i bold indikerer trialleliske SNPer. Navn i bold betyr triallelisk SNP og navn i rødt markerer en SNP som ble funnet med begge GBS-enzymer.

3.4.2 Tre SNPer med størst forskjell i allelfrekvens fra hvert GBS enzym

De tre SNPene fra hvert GBS-enzym med størst differanse i allelfrekvens mellom Tidlige og Sene (Tabell 9) hadde også alle så stor forskjell mellom tidlig og sen gruppe at alle hadde signifikans (P<0,01). To av disse hadde nabo-SNP med samme differanse og derfor er begge nevnt. I sammenheng med at jeg undersøker tre områder av hvert enzym teller jeg dem som en.

Tabell 6 SNPer med størst differanse mellom allelfrekvens i tidlige og sene.

Sted	GBS enzym	Differanse i allelfrekvens	Tassel Major/minor	*Major Allelfrekves Tidlige/Sene*	*SNP dybde (antall reads)*
scaf_766-7258	Pst1	0,58	T / C	*1,00 / 0,42*	*2195*
scaf_766-7250		0,57	A / G	*1,00 / 0,42*	*2196*
LG3-5423967	Apek1	0,55	T / C	0,85 / 0,30	*1129*
LG3-2673760	Apek1	0,53	T / C	0,36 / 0,89	*928*
scaf_119-13174	Apek1	0,49	C / T	*0,34 / 0,82*	*822*
LG7-6063939	Pst1	0,36	T / C	*1,00 / 0,64*	*4895*
scaf_350-136703	Pst1	0,35	- / A	*0,53 / 0,88*	*4704*
scaf_350-136724			G / -		*4703*

Høyeste differanse i allelfrekvens mellom tidlige og sene var på 0,58 i scaffold_766 posisjon 7258 og nabo-SNP på posisjon 7250 med omtrent identiske antall reads. De ble funnet med Pst1 (Tabell 6). Disse er homozygote for i den tidlige gruppen. Allelfrekvensen av det ene genet er tilnærmet 1 i den tidlige gruppen og 0,42 den sene gruppen (Tabell 6). SNPene ligger i en intron i et «Pentatricopeptide repeat» (PPR) gen. PPR og de andre nærmeste genene, histone acetyltransferase og retrotransposon (Tabell 7). Området ser ut til å ha genregulerende funksjon. SNPen ligger så nære enden av scaffoldet så vises bare ett gen i til i denne retningen (-1,6kb) på genombrowseren. Innenfor spennet på 25 kb ligger det 8 gener (tabell 7).

<u>Pst1</u>

Tabell 7 scaf_766-7258 funnet med pst1.Oversikt over nærliggende gener innen 12,5 kb

Avstand	ID	Funksjon
0, ligger i intron	gene29675	Pentatricopeptide repeat
+1 kb	gene29670	Pentatricopeptide repeat
-1,6 kb	gene29681	molecular_function unknown, (lokalisert i "H4/H2A histone acetyltransferase complex")
+4,3kb	gene29682	Retrotransposon protein
+9,6kb	gene29674	Sykdomsresistensrelatert protein (CC-NBS-LRR class)
+10,6kb	gene29678	Sykdomsresistensrelatert protein (CC-NBS-LRR class)
+10,8kb	gene29671	Sykdomsresistensrelatert protein (CC-NBS-LRR class)
+11,5kb	gene29684	Sykdomsresistensrelatert protein (CC-NBS-LRR class)

LG7-6063939 som ble plukket opp av begge enzymer men her representert av funn i Pst1. Det ligger 4 gener innenfor vinduet på +-12,5 kb (Tabell 8). SNPen ligger i siste basetriplett av siste exon i et DNA-primase gen ifølge genprediksjonen som vises i genombrowserene. De viser og at i referansegenomet er posisjon 6063939 på kromosom 7 basen T. De tidlige er homozygote for dette allelet. Den sene gruppen har allelfrekvens 0,64 for samme allel (Tabell 6). Når dette er en T, slik det og er fremstilt i referansegenomet så utgjør det STOP-kodon, TAG. Den alternative basen er C som koder for glutamin (CAG), og dermed vil translasjon fortsette istedenfor å stoppe her. Derfor er dette er SNP som antageligvis har direkte påvirkning på proteinproduktet. Hvis denne leserammen stemmer vil translasjon fortsette.

Tabell 8 LG7-6063939 Oversikt over nærliggende gener innen 12,5 kb

Avstand	ID	Funksjon
0, ligger i exon	gene9934	DNA primase, small subunit n=4
+1,2kb	gene9942	glycosyl hydrolase family protein 43
-3,8kb	gene9939	DNA primase, small subunit n=4
-6,9kb	gene9948	transmembran protein

SNPene på scaffold_350 posisjon 136703/136724 er intergenisk. Det ligger 4 gener i vinduet +-12,5 kb (tabell 9). Allelfrekvensen av det vanligste allelet i tidlige er 0,53 og i de sene 0,88 (Tabell 6).

Tabell 9 Scaffold_350-136703 Oversikt over nærliggende gener innen 12,5 kb

Avstand	ID	Funksjon
-2,3kb	gene18051	Cyclin D6
+4,3kb	gene18074	Bowman birk trypsin inhibitor
+6,3kb	gene18047	Ubiquinol-cytochrome C reductase iron-sulfur subunit
-11,7kb	gene18075	myb-like DNA-binding domain protein

<u>Apek1</u>

SNP på kromosom 3 posisjon 5423967 (LG3-5423967) ligger i exon på et actin-relatert protein. Innenfor vinduet +- 12,5 kb ligger det 4 gener (tabell 10). Forskjellen i allelfrekvens mellom tidlige og sene er 0,55. Frekvensen av det vanligste allelet var 0,85 i de tidlige og 0,30 i de sene (Tabell 6).

Tabell 10 LG3-5423967 Oversikt over nærliggende gener innen 12,5 kb

Avstand	ID	Funksjon
0, ligger i exon	gene6282	Actin-related protein Arp2/3 complex
-2,9kb	gene6186	transmembran protein
-5,8kb	gene6278	photosystem II core complex family psbY protein
-12,5kb	gene6246	S-adenosyl-L-methionine-dependent methyltransferases superfamily protein

SNP på kromosom 3 posisjon 2673760 (LG3-2673760) ligger i en exon i et gen som koder for «Transcription initiation factor TFIIE, beta subunit» Innenfor vinduet +- 12,5 kb ligger det 3 gener (Tabell 11). Her ligger det ett stressresponsivt alpha-beta barrel-domain protein og et metyl-CPG-bindene domene protein (MBD) som assosieres med genregulering. Forskjellen i allelfrekvens mellom de tidlige og sene er 0,55. Frekvensen av det vanligste allelet er 0,36 i de tidlige og 0,89 i de sene (Tabell 6).

Tabell 11 LG3-2673760 Oversikt over nærliggende gener innen 12,5 kb

Avstand	ID	Funksjon
0, ligger i exon	gene12978	Transcription initiation factor TFIIE, beta subunit
+3,7kb	gene12841	Stress responsive alpha-beta barrel domain protein
-3,7kb	gene12816	methyl-CPG-binding domain protein 13

SNP på scaffold_119 posisjon 13174 er intergenisk. Innenfor vinduet +- 12,5 kb ligger det 4 gener (Tabell 12). Forskjellen i allelfrekvens mellom tidlige og sene var 0,35. Det vanligste allelet hadde frekvens på 0,53 i de tidlige og 0,88 i de sene (Tabell 6). Området ser ut til å være knyttet til lipidmetabolisme. Gbrowse viste høyt GC innhold ~80% i området.

Tabell 12 scaf_119-13174 Oversikt over nærliggende gener innen 12,5 kb

Avstand	ID	Funksjon
+1,5kb	gene15722	patatin-like phospholipase
-4,6kb	gene33139	HXXXD-type acyl-transferase family protein
+8,9kb	gene33127	PATATIN-like protein 5
+10,5	gene33120	patatin-like phospholipase

3.4.3 Signifikant SNP som ble funnet med begge enzymer LG7-6063939

Den signifikante SNPen LG-7-6063939 ble funnet med begge enzymer (pst1 P<0,01)(apek1 P<0,05). SNPen hadde base T som major og C som minor over union av individene. I Pst1 var mange par av reads på denne SNPen fra de tidlige samplene filtrert bort av pearl-filteret på grunn av metning på 127 reads. Undersøkelse av råfilen viste at alle som var filtrert bort av pearlfilteret også var homozygote. Major allelfrekvens i de tidlige var 1. Total dybde for denne SNP med pst1 var 4895 som var litt over gjennomsnittet for SNP(P<0,05) funnet pst1 (Tabell 5). Dekningen av SNPen med Apek1 hadde dybde på 2878. Her var ingen prøver filtrert vekk på grunn av metning i reads. Major allelfrekvens i tidlige ble også estimert til 1. Blant de sene var frekvensen 0,64 målt med pst1 og 0,86 apek1 (0,22 i differanse). På grunn av forskjellen valgte jeg å sammenligne tagsekvenser fra området. Dessverre var ikke informasjon om antall reads tilgjengelig i 2088.sam fila så jeg visste ikke hvilke sekvenser som var representert i blant de signifikante SNPene og hvilke som inneholdt sekvenseringsfeil eller MAF<0,05. Derfor er det sannsynlig at enkelte av polymorfismene i sekvensene (Figur 7, 8 og 9) ikke er representative for populasjonen men av de jeg så på gikk det

igjen et mønster av at SNPen 6063939 lå både i og rett ved mulige RE-kuttseter. Først sammenlignet jeg to tagsekvenser fra hvert av enzymene mot hverandre (Figur 7). Den fra Apek1 var på 56 bp istedenfor 64 bp slik den fra pst1. Neste base på sekvensen fra Pst1 var en C og den utgjør siste base i kuttsete for Apek1. I figuren er det lagt til og kuttsete er merket i rødt (Figur 7). Eksempel på aligments jeg gjorde av tagsekvenser som strakk seg over SNPen (Figur 8, og 9) tyder det på at polymorfismen i LG6-6063939 og nabo-SNP 6063934 kan sammen føre til at det oppstår ett apek1 kuttsete fra ..34-..39 (GCTGC) eller ett Pst1 kuttsete fra ..36-..41 (CTGCAG). I pst1 har ikke SNPen «LG7-6063934» noen reads blant tidlige planter (Figur 10) mens med Apek1 er dem ikke detektert i det hele tatt. Eksempel i (figur 9) viser en to tagsekvenser som starter på samme sted, sekvens A på 48 bp, og sekvens B på 64 bp. Her er SNPen på posisjon 6063939 del av et kuttsete og er årsaken til at sekvensen bare er 48 bp. Restriksjonssetene rundt SNPen kan ha ført til uregelmessig dannelse av tagsekvenser avhengig av SNPen og kanskje derfor ulikt estimat i allelfrekvens ved bruk av de to enzymene.

Figur 7 Alignment av tagsekvens fra pst og apek1 for samme signifikante SNP som ble funnet LG7-6063939, SNP er merket med grønt men tagsekvensene fra de ulike enzymene representerer samme basevariant av denne SNPen. Denne tagsekvensen er forkortet til mindre enn 64bp på grunn av det antakeligvis er et apek1 kuttsete her. Markert i parentes med rød skrift. (GCWGC)

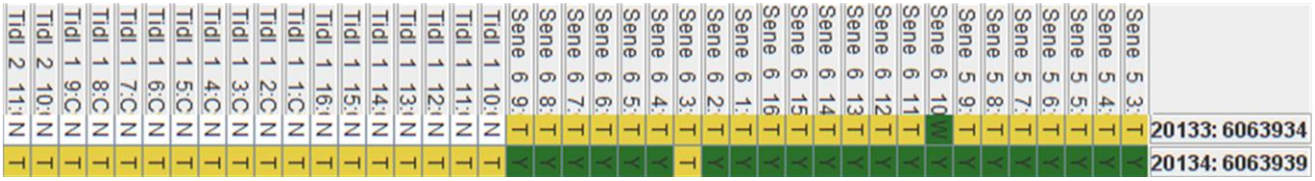
```
Apek1-6063892   2 TGCAGGTTCCAGTTTGCAGTCGAATTCGTGTGGTTCTGATATACTGTAGTGGCAG(C)   56 (L=56)
                  |||||.|||||||||||||||||||||.|||||||||||||||||||||||||||
Pst1-6063893    1 TGCAGATTCCAGTTTGCAGTCGAATTCGCGTGGTTCTGATATACTGTAGTGGCAG      55 (L=64)
```

Figur 8 Alignement av annen tagsekvens fra apek1 og samme fra pst1. Her viser en polymorfisme på posisjon 6063934 fra A til G (markert med grønt). I tillegg er det lagt til C i parentes som representerer den signifikante SNPen LG7-6063939 utfyller et restriksjonssete for apek1. Barcoden er antageligvis i denne enden og sekvensering har foregått nedover, mot starten på kromosomet i forhold til nummereringen i referansegenomet. Lengden er 64 i begge sekvenser

```
6063875(Apek1)   20 GCAGGTTCCAGTTTGCAGTCGAACTCGCGTGGTTCTGATATGCTG(C)   64  (L=64)
                    ||||.||||||||||||||||.||||||||||||||||||.|||
6063893(Pst1)     2 GCAGATTCCAGTTTGCAGTCGAATTCGCGTGGTTCTGATATACTG      46  (L=64)
```

Figur 9 Eksempel på alignment av to tagsekvenser fra Pst1 med samme nummerID fordi de starter samme sted. Type A sekvens er 48bp mens Type B sekvens er egentlig 64 bp lang. Sekvens B kuttes av Pst1 I polymorfismen C. Pst1 kuttsetet vises med rød skrift. Basen som er highligtet grønn er SNP-6063939. Nukleotid i parentes er ikke en del av tagssekvensen, men demonstrerer hvilken base som antakelig var tilstede og utgjorde et Pst1 kuttsete.

```
6063893-A        1 TGCAGGTTCCAGTTTGCAGTCGAATTCGCGTGGTTCTGATAAACTGCA(G)   48 (L=48)
                   |||||.|||||||||||||||||||||||||||||||||||||.|||||.|
6063893-B        1 TGCAGTTTCCAGTTTGCAGTCGAATTCGCGTGGTTCTGATATACTGTA      48 (L=64)
```



Figur 10 Skjermdump fra Tassel og visning av den filtrerte .vcf filen fra Pst1 av den signifikante SNPen LG7-6063939 som vises i bunn og over er nabo-SNP 6063934 (ikke sign). Det viser at den tidlige gruppen ikke hadde noen reads av 6063934. Dette er bare et utsnitt men mønsteret er det samme videre på de tidlige mot venstre og de sene mot høyre. «hvit N» betyr ingen reads. W betyr «A eller T», og Y betyr «C eller T»

3.5 Blastsøk etter feilkilder blant ualignede tagsekvenser

Det ble funnet flere tagsekvenser som stammer fra forurensinger i råfilen 2088.sam fila ved å manuelt finne ualignede tags og blastet disse på ncb og fikk relevante treff på andre organismer enn rødkløver (tabell 13).

Tabell 13 Noen sekvenser fra forurensing med gode hit på ncbi blast ble funnet i 2088.sam fila til Apek1 fra ualignede sekvenser.

Sekvens 1, med hit på *Metylobacterium Oryzae* 5e-08 CAGCGCCTGCGCCGCCAGGACGAAGTCGCCGGTCCCCGCGCGGTCGGAAATCGCCGAGACGTTG
Sekveks 2, megablast med hit på flere ulike fra *P. fluorescens* gruppen opptil 2e-22 CAGCGCGTGGTTCATCACCTTGAGCGCGGCATCGCCACGGCCCAGCAGGACTTCGTCCACCAGT
Sekvens 3, discontiguous megablast med hit på ulike *pseudomonas* opptil 4e-15 CAGCGCGTGGTCGATGGCGACACGGTGCGCCTGCGTGACGGCCGCAGTGTGCGGATGATTGGCC
Sekvens 4, discontiguous megablast med hit på biller 1e-09 CAGCGCCAGTCTCTGATGCAGTTTGATCTTCTTGACATGCCACTTTTGCTCGAGCGCGCGGTGG
Sevens 5, discontiguous megablast med hit på fugler opptil 4e-22 TGCAGGCGCTGAGAGCAGAGACCGGCGCGGGGTCCCCGGAGCCGGCTCGCCGCCCGCCGGGGGC
Sekvens 6, discontiguous megablast med hit på *Pantoea sp* opptil 1e-21 TGCAGGCGCTGACTCAGGCGGGCATCACGGCTGAGGCCGCGCCCCAGCCCGCACTTACCTCAGA
Sekvens 7, megablast med hit på menneske 5e-24 TGCAGTGGCACGATCTCAGCTAACTGCATCCTCTGCCTCCCGAGTTCAAGCGATTCTCCTGCCT
Sekvens 8, discontiguous megablast Mycobacterium sp. 1e-15 TGCAGTGGCACCACTTCCGATCACGACGACATTCTTGCCCGCGTAGTCCAGATCCTCCGGCCAG

4. Diskusjon

Feltforsøk

Gjennomsnittlig DTS (dager til strekning) blant de registrerte plantene med korrigering for sen registrering var bare 5,5 dager etter første registrering var fullført 23. oktober. Korrigeringsformelen for sen registrering var det beste tiltaket jeg hadde for å best mulig estimere tidspunkt for stengelstrekning. Et problem med denne var at selv om den var basert på en gjennomsnitt av 32 planter i varierende stadier var det i liten grad sammenheng mellom lengden på stengelen og daglig vekst de tidligere dagene og hastigheten varierte. Det var tydelig forskjell mellom de 52 tidligste og flesteparten av de andre 108 som allerede hadde startet strekning ved første registreringstidspunkt. Det var her korrigeringsformelen har hatt vært utslagsgivende. Det er synd at det er uvisst akkurat når tidligste plantene startet strekningsvekst i forhold til resten av plantene som ble godt kartlagt. Det var en tabbe starte registreringen såpass forsinket. Gruppene tidlige og sene som skulle karakteriseres genetisk var likevel betydelig adskilt fra hverandre i tidspunkt for stengelstrekning og de ytterste ekstremene i de tidligste ble uten tvil inkludert. At bare 12 av de 107 plantene som ikke hadde strukket seg innen 4.12.15 hadde begynt strekning 19.1.16 forteller at det var et passende tidspunkt å avslutte registreringen. Det ga en god representasjon av de seneste plantene uten disse. Det ville vært uforholdsmessig dyrt og tidssløsende å fortsette registreringene. De resterende plantene ble overvintret kaldt i drivhus og alle blomstret påfølgende sommer. Det kan tyde på at de hadde et vernaliseringskrav.

Bladene hadde begynt å bli bleke når plantene ble gjødslet for første gang (Bilde 2 vedlegg#). Det rettet seg raskt opp påvirket antakeligvis ikke differensieringen mellom tidlige og sene planter. Kanskje stimulert til smitte av *Rhizobium* bakterier. Det vil antakeligvis ha lite å si for forsøket ettersom det ble gjødslet regelmessig. Etterhvert som plantene vokste seg store ble vannbehovet større og det førte til at plantene til tider opplevde tørke selv med daglig vanning, men det var ikke så alvorlig at plantene tok synlig skade. Plantene kunne alternativt vært kuttet ned men det ville forstyrret observasjonene av andelen strekte skudd. Enveis Anova av om bordplassering påvirket DTS ga P-verdi på 0,347 så bordplassering hadde ikke signifikant betydning på DTS.

Delen av forsøket angående antall stengler i vegetativ fase og i strekning blir påvirket av usikkerheten av når de tidligste faktisk startet og det påvirker snittet av gruppens DTS og når 42 dager senere blir estimert til. Forholdet mellom strekte og vegetative skudd bekreftet at det var en sammenheng mellom strekningstidspunkt og andel skudd i strekning når målingene ble utført seks uker etter gjennomsnittlig tid for begynnende strekning. Den sene hadde høyere antall vegetative skudd og lavere andel skudd i strekning den tidlige gruppen. Forskjellen mellom tidlige og sene var større i antall vegetative skudd enn antall strekte skudd. Det kan indikere at plantene stor grad

slutter å produsere nye skudd etter påbegynt stengestrekning. Registrering av synlig blomst/knopp på dette tidspunktet viste at plantene som strakk seg sent også brukte lengere tid fra påbegynt strekning til synlig blomst enn plantene som strakk seg tidlig. Det kan være en sammenheng mellom enkelte gener som styrer utviklingstidene til disse to egenskapene.

Preparering av DNA

På grunn av 2 pipetteringsfeil fikk pool «sene1» 2 individer med omtrent 10% for mye DNA. Dette fører til en overrepresentasjon av disse to individene i poolen. Det var uheldig, men 10% er ikke så mye mer enn usikkerheten på kvantifiseringsmetodene (picogreen og qubit). Det er allerede en viss usikkerhet knyttet til kvantifiseringen. Konfidensvarians i triplettene målt i picogreen var i mange tilfeller over 10%.

GBS-data

Pst1 ga en høyere andel av signifikante SNPer enn apek1 i forhold til antall SNPer som ble undersøkt etter pearl filtreringen men Apek1 fant flere SNPer på flere steder. Med apek1 var tettheten av alle SNP funnet over 10 ganger høyere enn i med Pst1 etter «pearlfilteret» (Tabell 3). Kanskje det er påvirket av at Pst1 har større dybde (antall reads) per SNP enn Apek1 i dataen produsert fra Cornell (Vedlegg 6, Tabell 4 og 8). Det er Det kommer av at i Apek1 blir reads fordelt på flere fragmenter da den kutter oftere. En stor andel av SNPer i Pst1 blir filtrert bort på grunn av metning på 127 reads. Det kan være en ulempe at mange SNPer filtreres bort og da kan interessante SNPer ha bli tapt. Estimatet av allelfrekvens var i stor grad konsekvent mellom replikate av samme pool. Antallet SNPer som ble funnet samsvarer ikke helt med antall områder fordi det ligger noen i nære hverandre. Dette bør ikke påvirke fdr i sævlig grad fordi nærliggende SNPer bør være tilstede i samme grad også blant de usignifikante. Multi Dimensional Scaling plottene i hver av GBS-rapportene fra Genomic Diversity Facility, Cornell (vedlegg 6, Figur 3 og 5) kan tidlige og sene adskilles over en av aksene i både Pst1 og Apek1. Begge Det er en god indikasjon på at de to fenotypiske gruppene som ble laget også er genetisk forskjellige. Gruppen Tidlige_2 var plassert langt ut i begge plottene. Ingen andre var lengre unna noen annen gruppe. Kanskje denne poolen inneholdt en eller flere individer som stammer fra krysspollinering fra en annen populasjon eller kanskje det skyldes forurensing.

Distribusjon av SNPer

Flest SNPer ble funnet i LG7 og nest flest ble funnet i LG6. Færrest ble funnet i LG2 og LG3. Det er sannsynlig at det finnes flere områder i LG6 og LG7 som kontrollerer tidlighet av stengelstreking enn i de andre kromosomene. I *M. truncatula* er ble det i en studie på blomstringstid kartlagt flest på

kromosom 7 (Pierre et al. 2008) som er stor grad syntenisk til kromosom 6 i rødkløver (DeVega et al. 2015).

I 31 av SNPene med høyest signifikans (P<0,01) var en av gruppene homozygote for det ene allelet (allelfrekvens over 0,95) (Tabell 6). Disse kan være alleler som er dominante eller homozygot recessive. Mengden signifikante SNPer gjenspeiler antakeligvis det kompliserte samspillet av mange ulike gener som sammen styrer tidspunktet for stengelstrekning. I utgangspunktet er alle SNPene med signifikant forskjellig (P<0,05) allelfrekvens mellom gruppene av sent eller tidlig tidspunkt for stengelstrekning interessante. D

LG7-6063939 som ble funnet med begge enzymer

At SNP LG7-6063939 ble funnet i begge GBS run (Pst1 og Apek1) indikerer at metoden fungerer fordi samme område ble plukket opp av begge enzymer. Det er også en av SNPene med størst forskjell i allelfrekvens funnet med med Pst1 (Tabell 7). Men det estimert ulik allelfrekvens ved bruk av Pst1 og Apek1. Det kan være knyttet til at tilstedeværelsen av SNPen og nabo-SNP kan utgjøre et restriksjonssete i begge enzymer (Figur 7, 8, 9) Forskyvning i estimatet av allelfrekvens på grunn av SNP i restriksjonssetet er antakeligvis tilfeller gjennom hele prosessen og vil påvirke hvor nøyaktig allelfrekvensen estimeres. Dataene i .vcf filen som dette arbeidet er utført på gir ikke mulighet å se hvor på tagsekvenser SNP er hentet fra. Dersom den SNP ligger i restriksjonssetet vil den kanskje feilrepresenteres i antall reads. Samtidig amplifiseres kortere fragmenter i høyere grad enn lengere ved PCR. Dette kan også forskyve estimater av allelfrekvens ytterligere. Ulikheten har og forårsaket at SNPen får ulikt signifikansnivå i de to enzymene. Dette demonstrerer at det er en usikkerhet som kan gå igjen i alle SNPene som ble detektert. Selv om estimatene kan være unøyaktige ble LG7-6063939 funnet av begge enzymer og gjenspeiler en ulikhet som ble plukket opp. SNPen ligger i et STOP-kodon i et predikert gen som vil føre til at transkripsjonen fortsetter. Altså har den i seg selv en effekt. Den er homozygot i de tidlige og har stor forskjell i allelfrekvens. SNPen er god kandidat å undersøke videre. Området bør sekvenseres for å finne ut når det alternative STOP-kodonet kommer i allelet som er tilstede i mesteparten av de sene plantene.

Nærmere undersøkelse av SNPer som ser ut til å ligge i QTL for blomstringstid

Sammenligning av kromosomkartet laget i Mapchart over SNPer med signifikant ulik allelfrekvens mellom planter med tidlig tidspunkt for strekning og planter sent tidspunkt for strekning (P<0,05) med QTL linket til blomstringstid i rødkløver (Herrmann et al. 2006). Fire signifikante (P<0,05) SNPer så ut til å ligge i samme område som beskrevet i artikkelen. QTL på kromosomkartet med posisjon 37,6 cM og nabolokus C_E39/M59_380 ser ut til å være omtrent samme område som SNP LG5-

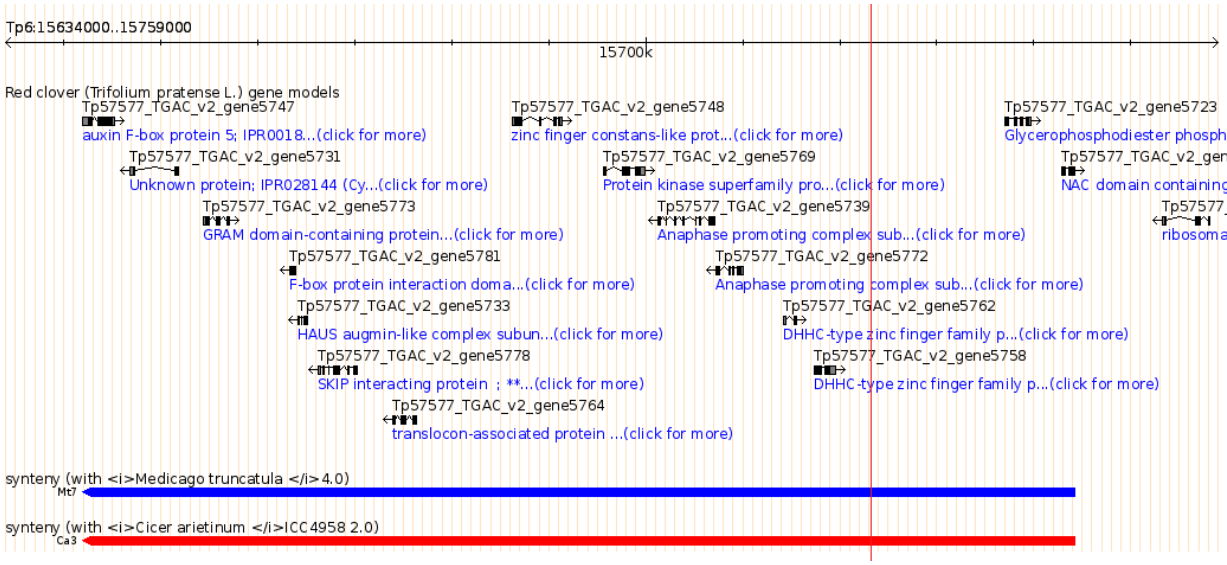
7766776 funnet med Apek1 (P<0,05) omtrent på midten av kromosom 5. SNPen er intergenisk men bare 109 bp fra nærmeste gen «RNA-dependent RNA polymerase 2; IPR000909».
QTL på 43,4 cM med nabolokus C_E39/M48_304 og markøren TPSSR23 ligger mot enden av kromosom 7. I samme område ser det ut til at LG7-25085103 funnet med Apek1 med (P<0,05) ligger. SNPen er lokalisert i et intron i gene-6902 Chalcone-flavanone isomerase family protein.
QTL på 44,8 cM med nabolokus V_P32/M15_65 ser ut til å gå over område der det er funnet to SNPer ble funnet. Trialleliske LG2-19944574 (P<0,05) funnet med Pst1. SNPen er intergenisk og nærmeste gen ene retning er +3,8 kb gene38325 «*homeobox-leucine zipper protein 17*» og andre veien – 12,2 kb gene38277 *Auxin efflux family carrier protein*. Også bialleliske LG2-20407052 ser ut til å ligge i samme QTL område. Den ligger i et intron på gene2660 *TCP-1/cpn60 chaperonin family protein*.

Nærmere undersøkelse av SNPer som ser ut til å ligge i QTL for vinteroverlevelse

QTL funnet av *(Klimenko et al. 2010)* knyttet til vinteroverlevelse ser ut til å ligge i samme område på kromosom 6 (45–78 cM) som noen SNP funnet med Apek1. Ettersom tidlighet av blomstring er knyttet til dårlig vinteroverlevelse (Choo 1984; Therrien & Smith 1960) er SNPene i dette området verdt å undersøke ekstra grundig. (Klimenko et al) fant tre QTL assosiert med vinteroverlevelse på nedre del av kromosom 6. I området så det ut som at det lå 4 SNPer funnet med Apek1 med signifikans (P<0,05). Følgende andre tre SNPer ble funnet i området: LG6-15026256 (P<0,05) Intron i gene3713 S-adenosylmethionine-dependent methyltransferase. Genene rundt ser ut til å være relatert til stresstolleranse/sykdomsresistens. LG6-15280022 (P<0,05) befinner seg i UTR på gene30361 *chalcone-flavanone isomerase family protein*,.
LG6-15723224 (P<0,01) er Intergenisk, SNPen ligger ved et *Constans-like* gen. Det beskrives i neste avsnitt om SNPer i område syntenisk med QTL assosiert med blomstringstid i *M. truncatula*.
LG6-16213525 (P<0,01) Nærmeste gen ligger +69 bp. (gene18669 heat shock protein 70) Genene rundt ser ut til å blant annet være relatert til stresstolleranse/sykdomsresistens. Nevner at SNP ligger 31,1 kb fra et «SCAR/WAVE familie»-gen som er assosiert til blant annet omorganisering av cellefunksjon. Det er bare putativt kartlagt tre *SCAR-like* gener i rødkløver. Homologer i *M. truncatula* er assosiert med rothårmorforlogi og rhizobium infeksjon av rothårene (Miyahara et al. 2010).

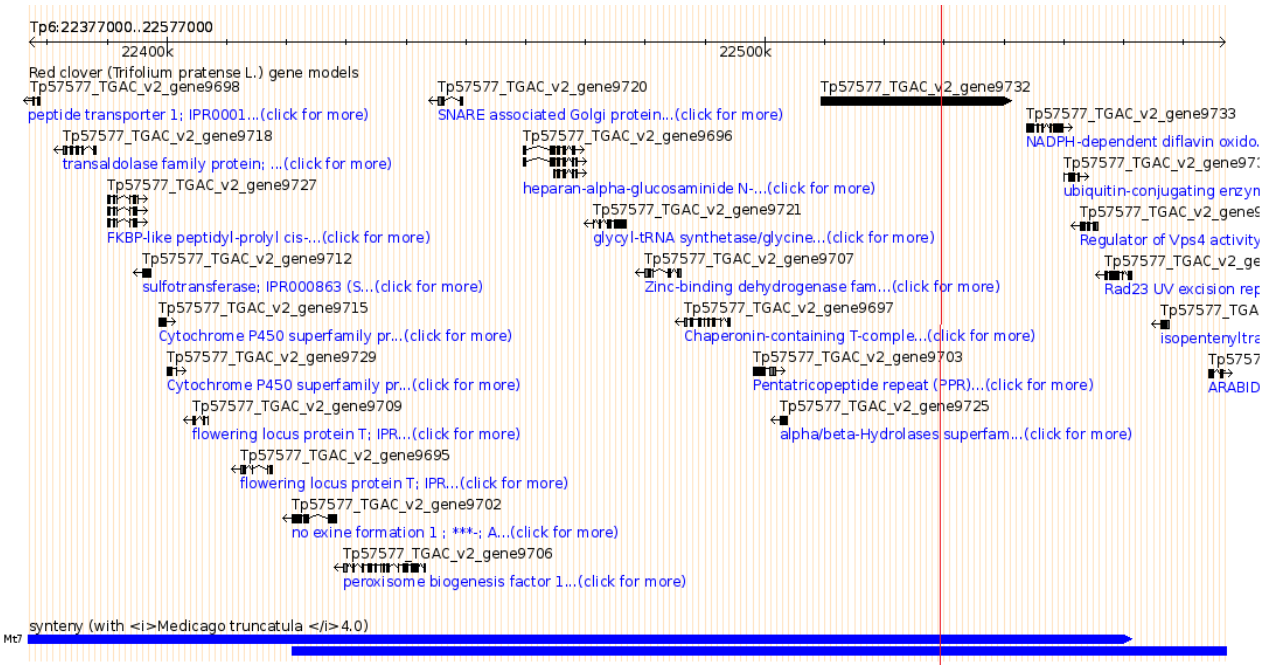
SNPer i område syntenisk med QTL assosiert med blomstringstid i *M. truncatula*

Det ble funnet en SNPer i QTL assosiert med blomstringstid i *M. truncatula* 57,2–58,1cM i kromosom 7 funnet av Pierre et al. (2008). Den ble funnet med Apek1 med signifikans (P<0,01). LG6-15723224 ligger i enden av en interessant cluster med gener. Det synteniske området med *M. truncatula* (og *C. arietinum*) strekker seg over 80,4 kb. Det nærmeste genet til SNPen -3,7 kb var et «*DHHC-type zinc finger familie protein*» med palmitoyl-trasferase aktivitet. Slike proteiner er lite beskrevet i planter, men utgjør viktige regulative funksjoner i eukaryote. Et annet DHHC-type zinc-finger protein funnet i *Arabidopsis thaliana* «At5g04270» er knyttet til regulering av skuddforgreining (J. Xiang et al 2010). Nedover i samme cluster av gener ligger det (-31,8 kb) «zinc-finger constans-like protein» (Figur 11). CONSTANS genfamilien er sterkt knyttet til blomstringstid og via fotoperiodiske signaler i *A. thaliana* (Putterill et al. 1995). I belgveksten og langdagsplanten *M. truncatula* er ikke kanskje ikke koblingen mellom blomstringstid og CONSTANS-like sterk (Weller & Ortega 2015; Wong et al. 2014). Gbrowse viser at området der jeg fant denne SNPen er syntenisk med et område på *M. truncatula sitt* kromosom 7. Her finnes og et «*zinc-finger constans-like*» gen, *Medtr7g083540/AC133780.7*. Dette er den samme som gen *CONSTANS-like* i QTL som styrer blomstringstid beskrevet av (Pierre et al. 2008; Pierre et al. 2011). På grunn av dette QTL-funnet ble dette genet i *M. truncatula* også undersøkt av Yeoh C.C. et al (2013) på den tidligblomstrende mutanten *spring1* men observerte ingen forskjell i ekspresjon av mRNA for Constans-like *Medtr7g083540* i forhold til villtypen. De fant derimot stor forskjell i ekspresjon av *flowering time locus t*. Det er foreslått at Constance kanskje ikke er like sterkt knyttet til daglengde som i andre planter (Wong et al. 2014). Området rundt SNP LG6-15723224 er interessant og bør undersøkes nærmere. Allelfrekvensen er estimert til 0,75 i de tidlige og 1 i det sene. Altså er dette allelet homozygot i gruppen med sen stengelstrekning.



Figur 11 Visning i Gbrowse av området rundt SNP LG6-15723224 (SNP er markert med rød linje) som er syntenisk med QTL i M. truncatula som inneholder genet CONSTANS (Pierre et al. 2008; Pierre et al. 2011). Intervallene på målestokken er 10 kb.

Avslutningsvis vil jeg nevne et annet gruppe gener i *M. truncatula* som påvirker blomstringstiden og er homologe til «*flowering locus T*» fra *Arabidopsis* (Laurie et al. 2011). LG6-22529694 (P<0,01) ligger 112 kb fra to «*flowering locus protein t*» gener i som er homologe med *MtFTb1* og *MtFTb2* og områdene er klassifisert som synteniske (Figur 12). Antakeligvis er ikke gene koblet til SNPen fordi det ligger tjuesju usignifikante SNPer mellom (men en i nabogenet no exine formation som var signifikant på 0,01 nivå i kun en retning med fst fordi en pool avvikte for mye). Det kan være at SNPen er koblet til et gen som regulerer *FLOWERING LOCUS protein T*. Laurie et al. (2011) så at *MtFTb2* transkripsjonsnivået økte i *M. truncatula* før blomstring og deretter sank omtrent når blomstene ble synlige. Genene ble uttrykt under lange dager, men ikke målbart under korte dager. Studien foreslår at variasjonen mellom FT-gener i en art kan danne grunnlag for lokale adapsjoner. Studie av Pierre et al. (2010) foreslår også at *MtFTb2* kan være et kandidatgen for kontroll av tidspunkt for blomstring. SNPen ligger i en foreslått gen-modell, gene9732 med lengde på nesten 32kb uten predikert genfamilie. Major allelfrekvens er 1 i gruppen med tidlig stengelstrekning og 0,884 i gruppen med sen stengelstrekning. Altså er alle de tidlige plantene homozygote for major allel. Selv om det er stor avstand mellom *FLOWERING LOCUS protein T* og SNPen som ble funnet, så er funnet interessant fordi genet er homologt til gener i andre planter med sterk tilknytning til blomstringstid.



Figur 12 Visning i Gbrowse av området rundt SNP LG6-22529694 (posisjonen er markert med rød linje) som er syntenisk med QTL i M. truncatula som inneholder genet MtFTb1 og MtFTb2 FLOWERING LOCUS-T homolog (Laurie et al. 2011) Intervallene på målestokken er 10 kb.

Blastsøk på ualignede tagsekvenser

Fra 2088.sam filen med tagsekvenser nblastet jeg noen sekvenser som ikke var alignet til referansegenomet for å se etter forurensinger (tabell 3). Blast treff på bille, fugl og menneske gjenspeiler at mikroskopiske forurensninger setter spor. Forurensing av andre organismer skal ikke ha påvirket resultatene. Eventuelle forurensinger er tilstede i svært liten grad. De fleste som ble funnet manuelt i 2088.sam filen kunne ha så lite som tre tagcount. Og ettersom tags har blitt alignet mot referansegenomet er ikke disse blitt med. Dersom tagsekvensen fra forurensningene stammer fra en konservert sekvens kunne den i teorien blitt tolket som en SNP dersom den samme konserverte regionen i rødkløver inneholdt RE sete. I det tilfellet ville likevel forurensningen blitt filtrert vekk da frekvensen trolig ville blitt langt under kravet for filtrering av minor allele frekvens. *Methylobacterium oryzum* er isolert fra stengelvev på risplante (M. J. Kwak M.J. et al. 2014). *Pantoea* er en annen familie bakterier matchet i blast som også lever på planter. *Pseudomonas* stammer antakeligvis fra jorda, *P.fluorescens* gruppen har flere stammer som kan ha positive interaksjoner med planter som å kunne gi planter beskyttelse mot enkelte patogene organismer.

Selv om bwa-alignmenten mot referansegenomet er god kan det alltid forekomme feil. Noen få sekvenser kan ha blitt feilalignet og andre kan ha blitt forkastet fordi de ikke ble alignet. BWA ble brukt til alignment hos Cornell. Metoden er god og mye brukt, men alle metoder har sin svakhet. Noen få SNPer kan ha blitt forkastet på grunn av at alignmenten kanskje ikke klarer å mappe noen prosent og ulike scoringsmatriser rangerer treffene ulikt selv om de er til samme området (Hatem et al. 2013).

Videre arbeid

Det ville vært en ide å sjekke SNPene mot andre individer karakterisert for blomstringstid. Det er ikke nødvendig å gjøre en GBS for å teste SNPene. Det kan designes primere til SNPene man ønsker å teste på en relativt enkel måte (Landegren et al. 1998). Det ville vært interessant å gjøre GBS med samme enzymer på andre sorter av rødkløver i et lignende eksperiment for å teste hypotesen om at disse SNPene er knyttet til tidspunktet for stengelstrekning og da kanskje inkludere den fenologiske gruppen som ikke blomstret det første året for å få bedre innblikk i hvilke av de samme eller nye SNPer som oppdages. Selv om forskjellen i allelfrekvens er konsis i «Lea» vil det være interessant å se hvilke SNPer som deles med sene varianter av rødkløver og som adaptert til et annet klima.

5. Konklusjon

Metoden så ut til å fungere og det ble funnet 218 SNP med signifikant forskjellig allelfrekvens (P<0,05) mellom gruppene med tidlig og sent tidspunkt for stengelstrekning. Det ble funnet flest signifikante SNPer på kromosom 6 og 7 med høyest konsentrasjon i nedre del på kromosom 6. Det ble oppdaget færrest signifikante SNPer i kromosom 2 og 3. Kunnskapen om områdene kan bidra til videre forskning og forståelse av hvordan tidspunkt for stengelstrekning kontrolleres i rødkløver.

Oppgaven har fått støtte fra forskningsprosjektet Agropro (2013-2017)

6. Referanser

Barnhart, S. K. & Rueber, D. (2013). Red Clover Variety Persistence Trial. *Iowa State Research Farm Progress Reports.* , Paper 2052.

Britten, E. J. (1963). Chromosome Numbers in the Genus Trifolium. *CYTOLOGIA*, 28 (4): 428-449.

Buxton, D. R., Hornstein, J. S., Wedin, W. F. & Marten, G. C. (1985). Forage Quality in Stratified Canopies of Alfalfa, Birdsfoot Trefoil, and Red Clover. *Crop Science*, 25 (2): 273-279.

Cassida, K. A., Griffin, T. S., Rodriguez, J., Patching, S. C., Hesterman, O. B. & Rustc, S. R. (1999). Protein Degradability and Forage Quality in Maturing Alfalfa, Red Clover, and Birdsfoot Trefoil. *Crop Science*, 40 (1): 209-215.

Choo, T. M. (1984). Association between growth habit and persistence in red clover. *Euphytica*, 33: 177-185.

Christie, B. R. & Martin, R. A. (1999). Selection for persistence in red clover. *Can. J. Plant Sci*, 79: 357-359.

DeVega, J. J., Ayling, S., Hegarty, M., Kudrna, D., Goicoechea, J. L., Ergon, Å., Rognli, O. A., Jones, C., Swain, M., Geurts, R., et al. (2015). Red clover (Trifolium pratense L.) draft genome provides a platform for trait improvement. *Scientific Reports*, 5.

Doležel, J., Bartoš, J., Voglmayr, H. & Greilhuber, J. (2003). Nuclear DNA content and genome size of trout and human. *Cytometry Part A*, 51A (2): 127-128.

Elshire, R. J., J. C. Glaubitz, Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S. & Mitchell, S. E. (2011). A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLoS ONE 6(5)*, e19379 (10.1371/journal.pone.0019379).

Ergon, Å. & Bakken, A. K. (2016). *Red clover traits under selection in mixtures with grasses versus pure stands*: Norwegian University of Life Sciences, Dept. of Plant Sciences, P.O.Box 5003, N-1431 Ås, Norway

Norwegian Institute of Bioeconomy Research, Dept. of Agricultural Technologies and System Analysis, Vinnavegen 38, 7512 Stjørdal, Norway. Upublisert manuskript.

Ergon, Å., Solem, S., Uhlen, A. K. & Bakken, A. K. (2016). Generative Development in Red Clover in Response to Temperature and Photoperiod. *Breeding in a World of Scarcity. Proceedings of the 2015 Meeting of the Section "Forage Crops and Amenity Grasses" of Eucarpia*: 243-269.

Glaubitz, J. C., Casstevens, T. M., Lu, F., Harriman, J., Elshire, R. J., Sun, Q., Buckler, E. S. & N. A. Tinker, E. (2013). TASSEL-GBS: A High Capacity Genotyping by Sequencing Analysis Pipeline. *10.1371/journal.pone.0090346*.

Graminor. *Lea, Rødkløver*. http://graminor.no/sorter/engvekster/rodklover/lea/: Bioforsk, Øst Løken (lest 23.01.2016).

Hatem, A., Bozdağ, D., Toland, A. E. & Çatalyürek, Ü. V. (2013). Benchmarking short sequence mapping tools. *BMC Bioinformatics*, 14 (184).

Hecht, V., Foucher, F., Ferra´ndiz, C., Macknight, R., Navarro, C., Morin, J., Vardy, M. E., Ellis, N., Beltra´n, J. P., Rameau, C., et al. (2005). Conservation of Arabidopsis Flowering Genes in Model Legumes. *American Society of Plant Biologists*, 137 (4): 1420-1434.

Hedrick, P. W. (2011). Genetics of Populations 4th ed.

Helgado´ttir, A., Larsen, A., Marum, P., Fritsen, H., Lindvall, E. & Miettinen, E. (2000). Prebreeding of Red Clover (Trifolium pratense L.) for Northern Areas. *Acta Agriculturae Scandinavica, Section B — Soil & Plant Science*, 50 (3): 187-190.

J. IŠTVÁNEK, M. JAROS, A. KŘENEK & ŘEPKOVÁ, J. (2014). GENOME ASSEMBLY AND ANNOTATION FOR RED CLOVER. *American Journal of Botany*, 101 (2): 327-337.

Jeff Glaubitz, R. E., Terry Casstevens, James Harriman, Ed Buckler. (2013). TASSEL 3 Genotyping by Sequencing (GBS) pipeline documentation.

Jones, T. W. A. (1974). THE EFFECT OF LEAF NUMBER ON THE SENSITIVITY OF RED CLOVER SEEDLINGS TO PHOTOPERIODIC INDUCTION. *Grass and Forage Science*, 29 (1): 25-28.

Klimenko, I., Razgulayeva, N., Gau, M., Okumura, K., Nakaya, A., Tabata, S., Kozlov, N. N. & Isobe, S. (2010). Mapping candidate QTLs related to plant persistency

in red clover. *Theor Appl Genet.* (120): 1253-1263.

Landegren, U., Nilsson, M. & Kwok, P. (1998). Reading Bits of Genetic Information: Methods for Single-Nucleotide Polymorphism Analysis. *Genome Res.*, 8: 769-776.

Laurie, R. E., Diwadkar, P., Jaudal, M., Zhang, L., Hecht, V., Wen, J., Tadege, M., Mysore, K. S., Putterill, J., Weller, J. L., et al. (2011). The Medicago FLOWERING LOCUS T Homolog, MtFTa1, Is a Key Regulator of Flowering Time. *Plant Physiol.*, 156 (4): 2207–2224.

Lunnan, T. N. L., Ås Inst. for Plantekultur). (1989). Effect of photoperiod, temperature and vernalization on floweing and growth in high latitude populations of red clover. *Norwegian Journal of Agricultural Sciences* (3): 201-210.

M. J. Kwak M.J., Haeyoung, J., Munusamy, M., Yi, L., Tong-Min, S., Tae, K. O. & Jihyun, F. K. (2014). Genome Information of Methylobacterium oryzae, a Plant-Probiotic Methylotroph in the Phyllosphere. *PLoS One*, 9.

Marum/Skog og Landskap. *Molstad rødkløver - medalje for frødyrking*: Skog og Landskap. Tilgjengelig fra: http://www.skogoglandskap.no/Artsbeskrivelser/moldstad/default_view (lest 23.01.2016).

Metzker, M. L. (2010). Sequencing technologies — the next generation. *Nature Reviews Genetics* 11: 31-46.

Miyahara, A., J. Richens, C. Starker, G. Morieri, L. Smith, S. Long, J. A. Downie & Oldroyd, G. E. D. (2010). Conservation in Function of a SCAR/WAVE Component During Infection Thread and Root Hair Growth in Medicago truncatula. *Molecular Plant-Microbe Interactions*, 23 (12): 1553-1562.

Montpetit, J. M. & Coulman, B. E. (1991). Responses to divergent selection for adventitious root growth in red clover (Trifolium pratense L.). *Euphytica*, 58 (119).

Pierre, J., Huguet, T., Barre, P., Huyghe, C. & Julier, B. (2008). Detection of QTLs for flowering date in three mapping populations of the model legume species Medicago truncatula. *Theor Appl Genet.* (117): 609-320.

Pierre, J. P., Bogard, M., Herrmann, D., Huyghe, C. & Julier, B. (2011). A CONSTANS-like gene candidate that could explain most of the genetic variation for flowering date in Medicago truncatula. *Molecular Breeding*, 28 (1): 25-35.

Putterill, J., Robson, F., Lee, K., Simon, R. & Coupland, G. (1995). The CONSTANS gene of arabidopsis promotes flowering and encodes a protein showing similarities to zinc finger transcription factors. *Cell*, 80 (6): 847-857.

Rafalski, J. (2010). Association genetics in crop improvement. *Current Opinion in Plant Biology*, 13 (2): 174-180.

S.R. Bowley, Taylor, N. L. & Dougherty, C. T. (1987). Photoperiodic response and heritability of the pre-flowering interval of two red clover (Trifolium pratense) populations. *Annals of Applied Biology*, 3: 455-461.

Sato, S., Isobe, S., Asamizu, E., Ohmido, N., Kataoka, R., Nakamura, Y., Kaneko, T., Sakurai, N., Okumura, K., Klimenko, I., et al. (2005). Comprehensive Structural Analysis of the Genome of Red Clover (Trifolium pratense L.) *DNA Research*, 12 (5): 301-364.

Sonah, H., Bastien, M., Iquira, E., Tardivel, A., Légaré, G., Boyle, B., Normandeau, É., Laroche, J., Larose, S., Jean, M., et al. (2013). An Improved Genotyping by Sequencing (GBS) Approach Offering Increased Versatility and Efficiency of SNP Discovery and Genotyping. *PLoS One*, 8(1).

Stephen Byrne, Adrian Czaban, Bruno Studer, Frank Panitz, Christian Bendixen & Asp, T. Genome Wide Allele Frequency Fingerprints (GWAFFs) of

Populations via Genotyping by Sequencing. *PLoS ONE*, 8 (3).

Therrien, H. P. & Smith, D. (1960). THE ASSOCIATION OF FLOWERING HABIT WITH WINTER SURVIVAL IN RED AND ALSIKE CLOVER DURING THE SEEDLING YEAR OF GROWTH. *Canadian Journal of Plant Science*, 40 (2): 335-344.

Undersander, D., Smith, R. R., Kelling, K., Doll, J., Worf, g., Wedberg, J. & Shaver, J. P. P. H. R. (1990). Red Clover Establishment, Manage ment and utilization. *University of Wisconsin--Extension*: 13.

Vince-Prue, D. (1975). Photoperiodism in plants. 444.

Vižintin, L., Javornik, B. & Bohanec, B. (2005). Genetic characterization of selected Trifolium species as revealed by nuclear DNA content and ITS rDNA region analysis. *Plant Science*, 170 (4): 859-866.

Weller, J. L. & Ortega, R. (2015). Genetic control of flowering time in legumes. *Frontiers in plant sciences*, 6 (207).

Wiersma, D. W. & Bolen, J. (2000). Red Clover Harvest Management. *Focus on Forage*, 3 (1): 1-2.

Wong, A. C. S., Hecht, V. F. G., Picard, K., Diwadkar, P., Laurie, R. E., Wen, J., Mysore, K., Macknight, R. C. & Weller, J. L. (2014). Isolation and functional analysis of CONSTANS-LIKE genes suggests that a central role for CONSTANS in flowering time control is not evolutionarily conserved in Medicago truncatula. *Frontiers in Plant Science*, 5 (486).

Wright, S. (1951). The genetical structure of populations. *Ann. Eugen*, 15: 323-354.

Vedlegg 1

Registreringer og statistikk på forsøket i drivhus

Boxplots viser 50% av gruppen i boksen (interquartile range box), resten ligger på den vertikale linjen, med unntak av outliere som markeres som *, streken over boksen er median og krysset på sirkelen er snitt.

Tabell 1 Oversikt over plantene som ble tatt videre til genetisk karakterisering. Tabellen oppgir estimatet for antall dager fra såing til strekningsvekst(DTS). Lengden på lengste stengel er oppgitt i cm og ble brukt i korrigeringsformelen for å estimere DTS. Antallet vegetative og strekte skudd registrert 6 uker etter gruppens gjennomsnittlige DTS

Tidlig gruppe			Antall skudd 6 uker etter gruppens snitt av DTS		**Sen gruppe**			Antall skudd 6 uker etter gruppens snitt av DTS	
PlanteID	DTS	Lengde ved registrering	Vegetative	Strekte	PlanteID	DTS	Lengde ved registrering	Vegetative	Strekte
517	23	35	0	3	**218**	70	2	7	2
308	24	34	0	3	**406**	70	2	8	4
451	24	33	0	3	**277**	71	3	8	3
231	26	35	1	4	**388**	71	3	2	3
100	29	28	4	3	**542**	71	3	1	2
213	29	29	6	1	**150**	72	2	7	5
217	30	26	2	3	**276**	72	2	6	5
133	31	29	1	6	**278**	72	4	3	5
378	31	25	1	4	**510**	72	2	10	2
476	31	27	2	4	**282**	73	3	6	5
473	32	24	2	3	**306**	73	3	4	3
547	32	24	2	3	**167**	74	2	5	3
500	33	27	2	4	**384**	74	2	4	5
302	34	22	2	5	**489**	74	2	5	3
438	34	23	0	4	**178**	76	2	6	3
112	36	21	2	5	**345**	76	2	6	3
368	36	19	3	3	**361**	76	2	4	5
625	36	19	3	3	**397**	76	2	6	4
48	37	19	2	5	**607**	76	2	4	3
198	37	18	2	5	**647**	76	2	3	2
422	37	19	3	5	**293**	79	3	7	3
365	38	17	2	4	**349**	79	7	6	4
200	39	17	1	5	**481**	79	3	6	4
360	39	16	4	3	**42**	80	2	8	4
457	39	17	2	5	**120**	80	2	7	2
529	39	16	2	4	**390**	80	2	6	1
650	39	17	2	3	**490**	80	2	5	4
138	40	15	2	5	**536**	80	2	6	2
312	41	15	1	5	**665**	80	2	5	3
531	41	15	2	4	**376**	81	4	10	3
13	42	13	1	3	**566**	81	8	7	5

16	42	13	2	5	**2**	82	7	3	3
131	42	13	3	4	**380**	82	7	5	4
172	42	12	2	4	**479**	82	16	6	3
191	42	13	3	3	**43**	83	2,5	5	6
546	42	13	1	4	**108**	83	15	8	3
17	43	13	2	4	**664**	83	5	2	1
121	43	13	3	4	**1**	84	4	7	5
233	43	13	2	4	**441**	84	14	4	4
286	43	13	2	6	**467**	85	3	5	2
309	43	11	3	6	**19**	86	2	5	2
324	43	11	2	4	**243**	88	9	4	5
332	43	12	2	4	**393**	88	9	4	4
347	43	13	0	6	**619**	91	5	8	3
446	43	12	2	4	**637**	91	6	4	3
507	43	11	3	4	**26**	92	4	8	3
514	43	15	1	5	**560**	92	4	3	2
155	44	10	2	6	**562**	92	4	6	4
281	44	11	2	4	**609**	92	4	6	3
313	44	11	1	4	**494**	93	3	5	1
472	44	10	1	5	**307**	94	2	10	4
550	44	10	2	3	**672**	94	2	2	5

Tabell 2 Datoer det ble registrert strekning

22.10.2015	01.11.2015	14.11.2015
23.10.2015	03.11.2015	16.11.2015
25.10.2015	05.11.2015	20.11.2015
26.10.2015	08.11.2015	23.11.2015
27.10.2015	10.11.2015	26.11.2015
30.10.2015	15.11.2015	04.12.2015

Tabell 3 Grunnlaget for utregning av korrigeringsformelen, EstDagligVekst. Etterregistreringen av disse plantene ble gjort 12.11.15

Tidsrom og vekst	cm/dag	dager	Plante #
7 dager 3 til 6 cm	0,42	7	252
4 dager 2 til 4 cm	0,5	4	118
2 dager 3 til 4 cm	0,5	2	105
2 dager 2 til 3 cm	0,5	2	214
7 dager 3 til 7 cm	0,57	7	254
9 dager 2 til 8 cm	0,66	9	123
7 dager 2 til 5 cm	0,71	7	228
4 dager 2 til 5 cm	0,75	4	203
13 dager 2 til 13 cm	0,85	13	144
13 dager 2 til 13 cm	0,85	13	171
13 dager 2 til 14 cm	0,92	13	220

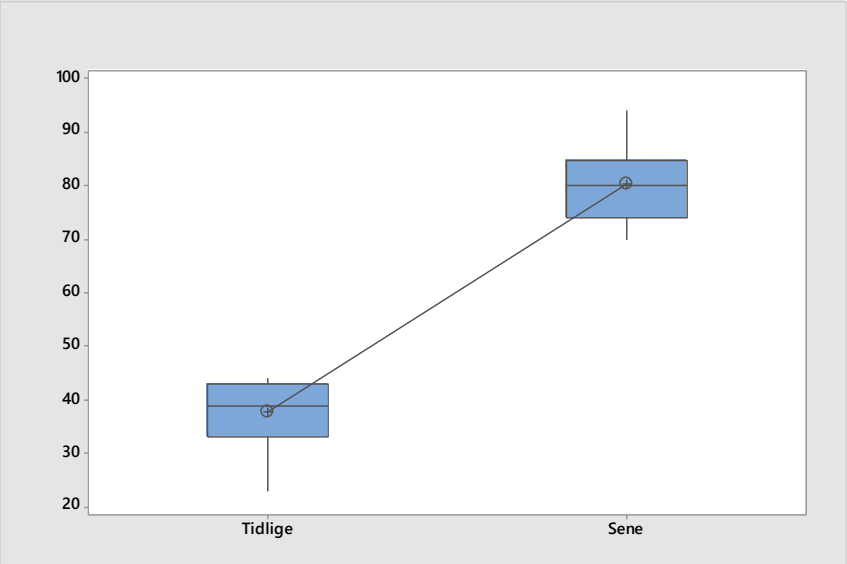
9 dager 2 til 11 cm		1	4	190
4 dager 2 til 6 cm		1	2	230
2 dager 2 til 4 cm		1	9	236
13 dager 3 til 16 cm		1	13	156
11 dager 3 til 15 cm		1,09	11	186
13 dager 3 til 18 cm		1,15	13	222
11 dager 3 til 16 cm		1,18	11	189
13 dager 3 til 20 cm		1,3	13	193
13 dager 3 til 20 cm		1,3	13	195
9 dager 2 til 14 cm		1,33	15	110
15 dager 2 til 22cm		1,33	9	212
9 dager 4 til 17 cm		1,44	13	221
4 dager 2 til 8 cm		1,5	4	188
13 dager 3 til 23cm		1,53	13	141
9 dager 2 til 17cm		1,66	9	115
13 dager 2 til 24 cm		1,69	13	176
11 dager 3 til 22cm		1,72	11	256
4 dager 3 til 10 cm		1,75	4	209
9 dager 3 til 19 cm		1,77	9	122
11 dager 3 til 24 cm		1,9	11	223
17 dager 4 til 47 cm		2,52	17	106
32 stk	snitt cm/dag	stdev	var	
	1,1684375	0,39959	0,24413619	

Tabell 4 Styrke på tilleggs-lyset per bord i µMol-1m2s, snitt av 3 målinger på hvert bord målt 30 cm fra bordoverflaten.

	Bord 1	Bord 2	Bord 3	Bord 4	Bord 5	Bord 6	Bord 7
µMol m2s1	86,67	121,00	116,67	102,67	100,00	62,67	51,33

Tabell 5 Toveis t-test på DTS mellom tidlige og sene planter

Two-sample T-test Tidlige; Sene (korrigerte verdier for sen registrering)			
Gruppe	Snitt	StAvik	StFeil snitt
Tidlige n=52	37,73	6,03	0,84
Sene n=52	80,52	7,22	1,0
P verdi = 0,000		T-verdi = 32,81	



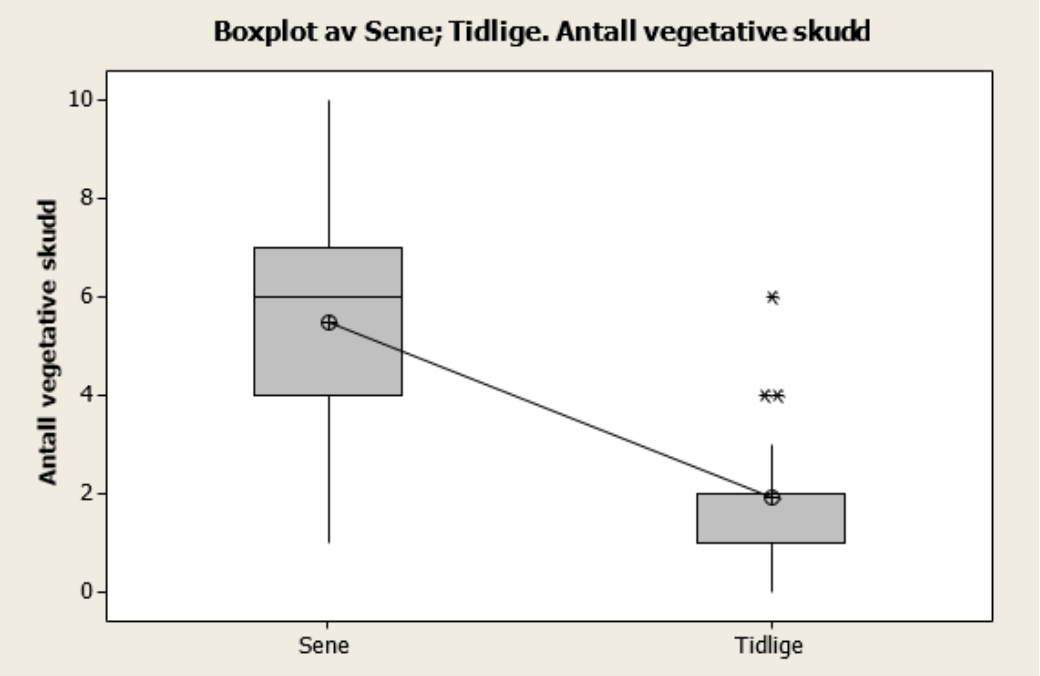
Figur 1 Boxplot Dager til strekning, DTS i Tidlig gruppe tv. og sen gruppe th., hver på 52 stk (med korrigerte DTS verdier ved sen registrering)

Tabell 6 Antall skudd i strekning og vegetativ vekst 6 uker etter gjennomsnittlig tidspunkt for strekning innenfor gruppen tidlige eller sene n=52.

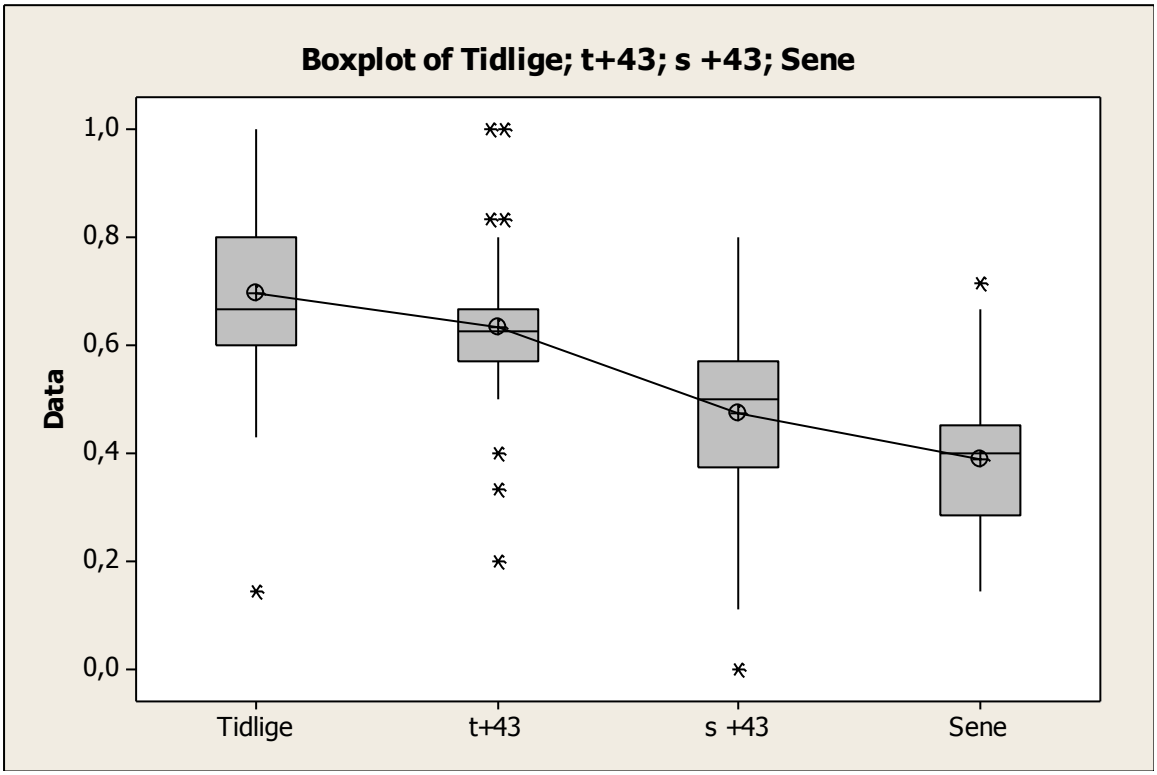
Tidlige			
	Veg.	Strek.	strek/totalt antall
Snitt	1,92	4,10	0,695
St.dev	1,10	1,03	0,159
Sene			
Snitt	5,50	3,33	0,388
St.dev	2,09	1,15	0,127
Tidlig vs sene	P<0,0001	P=0,001	P<0,0001
T-test	T=10,92	T=3,59	T=10,88

Tabell 7 Enveis Anova av alle bord på DTS for å se om romplassering hadde påvirkning på strekningstidspunkt. Kun planter med strekning innen 4.12.15 er med, 555stk. Forkrøplete og kastede planter er ikke med.

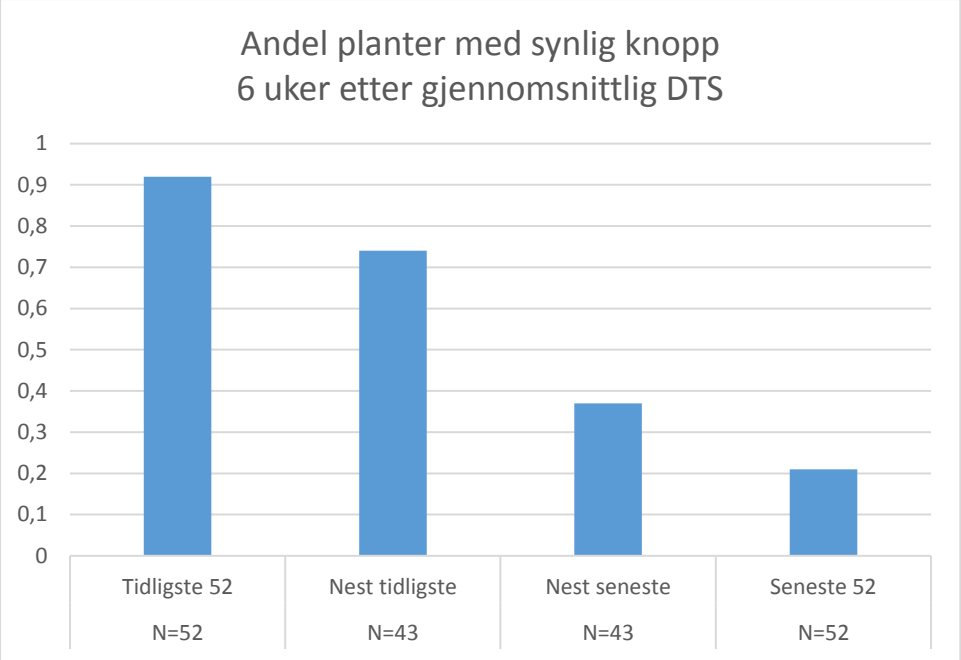
	Frihetsgrader	Sum squares	Mean squares	F	P
Bord	6	880	147	1,12	0,347
Residuals	548	71569	131		
Total	554	72449			
	S= 11,43	R-sq=1,22%	R-sq (adj) = 0,13%		



Figur 4 Antallet vegetative skudd i den sene og tidlige gruppen vist i boxplott.



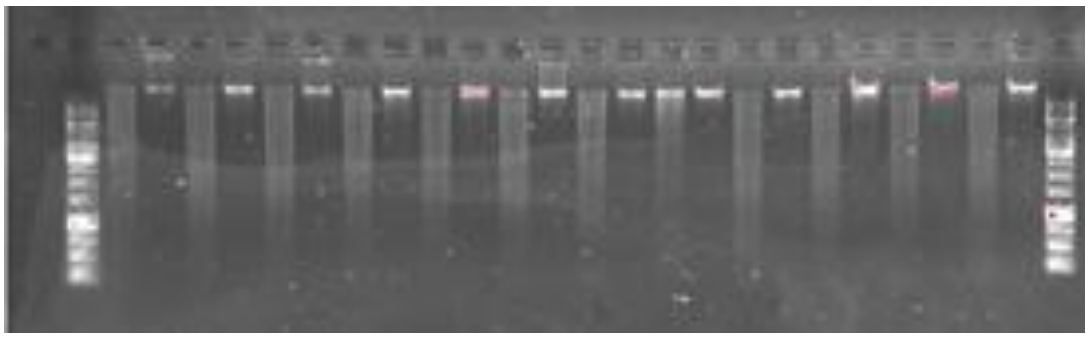
Figur 5 Boxplot av andel skudd i strekning 6 uker etter snitt av gruppenes DTS. Her er alle de 95 tidligste og 95 seneste fremstilt. Tidlige og Sene er gruppene som ble genetisk karakterisert. t+43 (Tidlige) og s+43 (Sene) er 43 stk planter som akkurat var for trege eller sene i strekning til å bli med i gruppene tidlige eller sene.



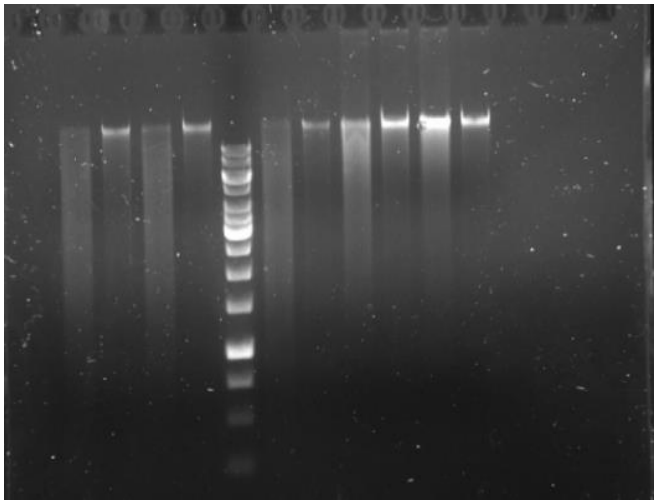
Figur 6 Figuren viser hvor stor andel av plantene som hadde blomst/synlig knopp 6 uker etter snitt av gruppenes DTS. Her er alle de 95 tidligste og 95 plantene fremstilt. Gruppene som ble genetisk karakterisert vises helt til venstre (tidlige) og helt til venstre (sene) planter som akkurat var for trege eller sene i strekning til å bli med i gruppene tidlige eller sene er vist imellom.

Vedlegg 2

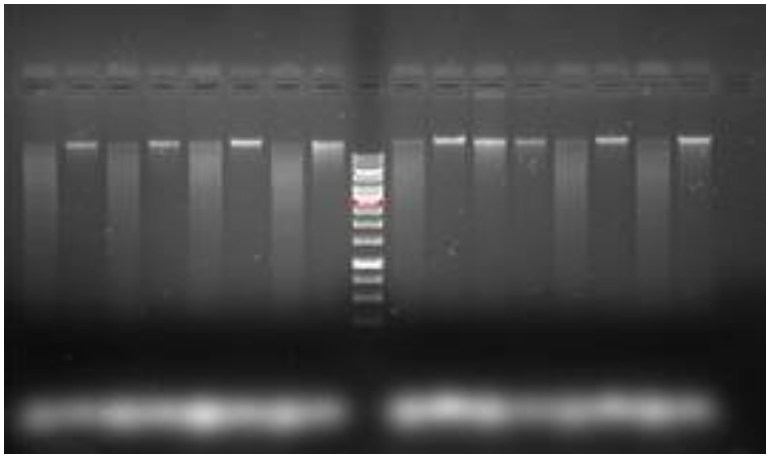
Testkutting med HindIII



Figur 1 Testet kuttbarhet av isolert DNA med HindIII. Par av den samme prøven kuttet(1) og ukuttet(1) står ved siden av hverandre. Tre prøver fra Tidlige planter første eluering (Qiagen 96 plant kit) 16, 48, 625 tv. + 2 prøver fra Qiagen Dneasy plant minikit 282, 397



Figur 2 Testet kuttbarhet av isolert DNA med HindIII. Prøvene er av tidlige og sene planter fra 2. eluering (Qiagen 96 plant kit). Par av den samme prøven kuttet(1) og ukuttet(1) står ved siden av hverandre.



Figur 3 Testet kuttbarheten med HindIII på prøver isolert med DNA med Dneasy Plant minikit (QIAGEN). Par av den samme prøven kuttet(1) og ukuttet(1) står ved siden av hverandre.

Tabell 1 Blandingsforhold for kuttingsreaskjonen. µl væske per prøve, ganges opp med antall prøver

Ingrediens	**Per prøve (mikroliter)**
10x NEBuffer2	1,5
20 U/mikroliter HindIII	0,05 (tilsvarer 1U)
Vann	Y (6)
Tilsettes enkeltrør etter fordeling av mastermix:	
DNA	X (tilsvarende 100 ng)
Vann	Z
Totalt per rør	15 µl

Vedlegg 3

Prøveplassering i plater for GBS

Tabell 1 Plassering av prøver fra de seks poolene over 96-platen til GBS med Apek1 som enzym. Nummering av DNA-pool 1-6 med navn som forklarer hvilken kategori den tilhører. Tidlige(T) planter 1, 2, 3 og Sene (S) planter 4, 5, 6. Svart plass angir blank.

Apek1

	1	2	3	4	5	6	7	8	9	10	11	12
A	TP1	TP1	TP1	TP1	TP1	TP1	TP1	TP1	TP1	TP1	TP1	TP1
B	TP1	TP1	TP1	SP4	SP4	SP4	SP4	SP4	SP4	SP4	SP4	SP4
C	SP4	SP4	SP4	SP4	SP4	SP4	SP4	TP2	TP2	TP2	TP2	TP2
D	TP2	TP2	TP2	TP2	TP2	TP2	TP2	TP2	TP2	TP2	TP2	SP5
E		SP5	SP5	SP5	SP5	SP5	SP5	SP5	SP5	SP5	SP5	SP5
F	SP5	SP5	SP5	SP5	TP3	TP3	TP3	TP3	TP3	TP3	TP3	TP3
G	TP3	TP3	TP3	TP3	TP3	TP3	TP3	TP3	SP6	SP6	SP6	SP6
H	SP6	SP6	SP6	SP6	SP6	SP6	SP6	SP6	SP6	SP6	SP6	SP6

Tabell 2 Plassering av prøver fra de seks poolene over 96-platen til GBS med Apek1 som enzym. Nummering av DNA-pool 1-6 med navn som forklarer hvilken kategori den tilhører. Tidlige(T) planter 1, 2, 3 og Sene (S) planter 4, 5, 6. Svart plass angir blank.

Pst1

	1	2	3	4	5	6	7	8	9	10	11	12
A	TP1	TP1	TP1	TP1	TP1	TP1	TP1	TP1	TP1	TP1	TP1	TP1
B	TP1	TP1	TP1	TP1	SP4	SP4	SP4	SP4	SP4	SP4	SP4	SP4
C	SP4	SP4	SP4	SP4	SP4	SP4	SP4	TP2	TP2	TP2	TP2	TP2
D	TP2	TP2	TP2	TP2	TP2	TP2	TP2	TP2	TP2	TP2	TP2	SP5
E	SP5	SP5	SP5	SP5	SP5	SP5	SP5	SP5	SP5	SP5	SP5	SP5
F	SP5	SP5	SP5	TP3	TP3	TP3	TP3	TP3	TP3	TP3	TP3	TP3
G	TP3	TP3	TP3	TP3	TP3	TP3	TP3	SP6	SP6	SP6		SP6
H	SP6	SP6	SP6	SP6	SP6	SP6	SP6	SP6	SP6	SP6	SP6	SP6

Vedlegg 4

Signifikante SNPer (P<0,05)

Tabell 1 *SNPer med ulik allelfrekvens i planter med sen stengelstrekning og planter med tidlig stengelstrekning (P<0,01). Informasjon om lokasjon, synteni er hentet fra Jbrowse og Gbrowse. Major refererer til det som ble scoret som major allel av Tassel som svarer til hvilket allel som hadde flest reads over alle prøvene totalt. Venstre kolonne viser posisjonen til SNPen og i hvilket kromosom eller scaffold den befinner seg. LG står for «linkage group#» som er synonymt med kromosomnummer.*

Mappet posisjon på ref.genomet kromosom eller scaffold	**GBS Enzym**	**SNP Major/ Minor**	**Lokasjon**	**Syntenisk område i *Medicago truncatula (M)* eller *Cicer arietinum (C)***	**Snitt av frekvens Major allel (og 1. minor i triallele)**		**Dybde**
					Tidlig gruppe	**Sen gruppe**	
LG1-2331127	Pst1	G / A	Exon i gen27071 (pumilio-family RNA-binding repeatprotein)	M, C	0,75	0,98	2084
LG2-5298235	Pst1	A / G	Repetetivt TA-område Ikke i gen, ca -3,5kb til nærmeste	M, C	0,61	0,88	4895
LG3-2673758	Apek1	G / A	Exon i gene12978 Transcripsjon initieringsfaktor TFIIE, beta subunit	M, C	0,91	0,63	928
LG3-2673760		T / C			0,36	0,89	928
LG3-5423967	Apek1	T / C	Exon i gene6282 Actin-related protein Arp2/3 complex, subunit Arp3	M, C	0,85	0,30	1129
LG3-30484503	Pst1	T / indel	Intron i gene9044: double-stranded RNA-binding motif protein	-	0,98	0,64	5848
LG3-30484546		-/T/A			0,98 (0,01)	0,64 (0,36)	5848
LG4-5053247	Apek1	G / A	Exon i gen24777 ATP binding/protein serine/threonine kinase [Glycine max]	M	0,49	0,79	767
LG5-906350	Apek1	T / C	Exon i gen27326 delta-1-pyrroline-5-carboxylate synthetase	M	0,51	0,85	1368
LG5-1283650	Apek1	C / A	Exon i gen7910 F-box/RNI-like superfamily protein	M, C	0,97	0,78	757
LG5-1283651		A / C					
LG5-1283653		A / T					
LG6-4207656	Apek1	C / T	Intergenisk ca -7 kb til nærmeste gen S-ribonuclease n=5	C	0,60	0,96	900
LG6- 10343986	Apek1	C / G	I exon i gen20073 subtilisin-like serine protease	M, C	0,95	0,78	4001
LG6-12647432	Apek1	C / T	Intergenisk, nærmeste ligger ca -2,5kb gene1075 GDSL-like Lipase/Acylhydrolase superfamily protein	M	0,98	0,80	3204
LG6-12647447		C / T			0,98	0,80	

LG6-15723224	Apek1	A / G	Intergenisk, nærmeste ligger ca -3,7kb gene5778 DHHC-type zinc finger family protein	M, C QTL (Pierre et al. 2008)	0,75	1,00	2703
LG6-16213525	Apek1	C / G	Intergenisk, men nærmeste gen er bare +70bp gene18669 Heat shock protein 70	M, C	0,66	0,36	2039
LG6-20401679	Apek1	G / C	Intron i gen10187 myb domain protein 118	M, C	0,62	0,89	780
LG6-20401706		A / C					
LG6-20871716	Apek1	G / A	Integenisk, nærmeste gen er ca -0,4 kb gene33586 ATP-dependent zinc metalloprotease FTSH protein	M, C	0,77	0,98	1349
LG6-21093073	Apek1	T / A	Exon i gen15632 ENTH/VHS family protein	M	0,67	0,34	1680
LG6-21093208		G / A			0,72	0,96	1778
LG6-21393289	Pst1	T / A	Repetivt T område. (ca 16bp langt) Intergenisk, ca 1kb til nærmeste gen	-	1	0,83	5511
LG6-21393294		-/T/C			0,30 (0,15)	0,33 (0,53)	7951
LG6-21881050	Apek1	T / A	Intergenisk, nærmeste gen -2,5kb HXXXD-type acyl-transferase family protein	M, C	0,40	0,86	1532
LG6-22529694	Apek1	C / G	Intergenisk, nærmeste gen +14,2kb	M, C (Laurie et al. 2011)	0,88	1	4608
LG7- 6603333	Apek1	A/Indel	Intergenisk, nærmeste gen +0,9 kb gene1361 beta-1,4-xylosyltransferase,	M, C	0,31	0,74	1380
LG7- 6603349		T / C			0,77	0,33	2689
LG7- 6603360		A / G			0,77	0,31	1380
LG7-6063939	Apek1 (P<0,05)	T / C	Exon i enden av gen9934 Lille subenhet på dna primase.	-	1,00	0,86	2878
	Pst1				1	0,64	4895
LG7-8706752	Apek1	A/Indel	Intron i gen39014 trafficking protein particle complex subunit-like protein	M, C	0,95	0,73	973
LG7-8706753		A/Indel					
LG7-8924269	Pst1	Indel/T	Intron i gen34550 Ukjent protein	M, C	0,85	1,00	6458
LG7-8924289		G/C			0,85	1,00	
LG7-8924291		C/Indel			0,85	1,00	
LG7-11248005	Pst1	G/A	Intergenisk, ca -5,8bp til nærmeste, gene3141 Bifunctional inhibitor/lipid-transfer protein/seed storage 2S albumin superfamily protein	M, C	0,97	0,79	5830
LG7-14523935	Pst1	A/G	UTR i gene11523: FASCICLIN-like arabinogalactan-protein 11	M, C	1,00	0,67	3855
LG7-14523937		G/A			1,00	0,67	
LG7-15175628	Apek1	A / G	Intergenisk, nærmeste gen ligger +0,4 kb Ribonuclease P protein subunit og -4,2 kb CCT motif family protein	M, C	0,79	1,00	4672
LG7-15789143	Apek1	T / - / C	Intergenisk, nærmeste gen ligger +0,4 kb gene30210 CW7 Ukjent funksjon	M, C	0,77	0,92	1290

LG7-17998574	Apek1	G /T / A	Intron i geen13079 annexin D8	M, C	0,36 (0,49)	0,66 (0,17)	1973
Scaffold82-515442	Apek1	C / T	Exon i gene3293 myosin-like protein	M, C	0,39	0,86	3862
Scaffold119-13174	Apek1	C / T	Intergenisk, nærmeste gen ligger +1,5 kb gene33148 patatin-like phospholipase	-	0,34	0,82	822
Scaffold152-121219	Apek1	C / A	Intergenisk nærmeste gener er +1,1 kb og – 1,5 kb serpin like protein	M, C	0,82	0,55	8216
Scaffold267-267337	Apek1	G / A	Intron i gene16357 DNA-directed RNA polymerase family protein	-	1	0,76	1323
Scaffold320-70770	Apek1	T / A	Intergenisk, nærmeste gen er +5 kb receptor like kinase 1	M	1	0,86	963
Scaffold350-136703	Pst1	- / A	Intergenisk, nærmeste gen er ca -2,3kb	M, C	*0,53*	*0,88*	4703
Scaffold350-136724		G / -					4704
Scaffold442-72281	Apek1	A / C	Exon i gene14066 Pentacropeptide repeat	M, C	1,00	0,85	1635
Scaffold512-170959	Apek1	C / T	Exon i gene8770 Pentacropeptide repeat	M	0,83	1,00	865
Scaffold700-31116	Apek1	T / G	Exon i gene3178 P-loop nucleoside triphosphate hydrolase superfamily protein	C	1,00	0,86	2606
Scaffold702-112119	Apek1	G / - / T	Intron i gene31753 Werner syndrome ATP dependent helicase-like protein n	M, C	0,88 (0,07)	1,00 (0)	1613
Scaffold766-7250	Pst1	A/G	Intron i gene29675 Pentatricopeptide repeat	-	1,00	0,42	2195
Scaffold766-7258		T / C			1,00	0,42	2196
Scaffold893-28441	Apek1	A / G	Exon i gene28132 pyruvate orthophosphate dikinase	M, C	1,00	0,84	777
Scaffold956-7118	Apek1	A / G	Intergenisk, nærmeste gen er -2,8 kb, Ankyrin repeat family protein	-	0,86	0,62	2505
Scaffold956-7119		T / G					
Scaffold2623-2286	Apek1	G/ C/ A	Ingen kjente genmodeller i scaffoldet	M, C	0,09	0,41	1602
Scaffold2856-2938	Apek1	G / A	Ingen kjente genmodeller i scaffoldet	-	0,81	1,00	2445
Scaffold6992-1002	Apek1	G / A	Exon i gene382 Ubiquitin thioesterase otubain-like n	-	0,76	0,47	1329
Scaffold16044-479	Apek1	C / T / A	Ingen kjente genmodeller i scaffoldet	M, C	0,32	0,48	1108
Scaffold28709-135	Apek1	C / A	Exon i gene39834 Disease resistance protein (CC-NBS-LRR class) family	-	0,83	0,97	1131

Scaffold36286-233	Apek1	C / T	Ingen kjente genmodeller i scaffoldet	-	0,91	0,65	1384

Tabell 2 Bialleliske SNP funnet med pst1 med signifikant ulik allelfrekvens mellom gruppene karakteriser for tidlig stengelstrekning og sen stengelstrekning.(P<0,05). Posisjon på kromosomet er anngitt til venstre. Major / Minor er bestemt ut fra til hvilket allel som hadde flest reads i .vcf råfilen. Gjennomsnittlig allelfrekvens innen gruppene og differansen mellom snittene er vist under differanse

		SNP	Allelfrekvens Major				
Kromosom	Posisjon	Maj/ Min	Sene	Tidlige	Differanse	Dybde	Signifikansnivå
LG1	2331127	C/ T	0,98	0,75	0,23	2084	(P<0,01)
LG1	5349954	G/C	1,00	0,90	0,10	4172	(P<0,05)
LG1	12703423	G/A	1,00	0,83	0,17	3245	(P<0,05)
LG2	5298235	A/G	0,88	0,60	0,27	4895	(P<0,01)
LG2	19944596	C/T	0,54	0,84	0,30	6598	(P<0,05)
LG2	28426502	T/A	0,96	0,81	0,15	1548	(P<0,05)
LG3	30484503	T/-	0,64	0,98	0,34	5848	(P<0,01)
LG4	589451	A/G	0,92	0,73	0,20	1813	(P<0,05)
LG4	2409656	C/A	1,00	0,89	0,11	2687	(P<0,05)
LG4	8291691	T/A	0,99	0,77	0,21	3012	(P<0,05)
LG4	9554592	T/A	0,90	1,00	0,10	4327	(P<0,05)
LG4	9554609	C/T	0,90	1,00	0,10	4327	(P<0,05)
LG5	14810099	C/G	0,50	0,79	0,28	5193	(P<0,05)
LG6	9425884	G/C	0,89	0,99	0,10	4466	(P<0,05)
LG6	9622227	C/A	1,00	0,89	0,11	1607	(P<0,05)
LG6	11606887	C/T	0,81	0,95	0,14	4809	(P<0,05)
LG6	20354371	-/T	0,57	0,88	0,30	4387	(P<0,05)
LG6	21393289	T/A	0,83	1,00	0,17	5511	(P<0,01)
LG7	6063939	T/C	0,64	1,00	0,36	4895	(P<0,01)
LG7	8924269	-/C	1,00	0,84	0,15	6458	(P<0,01)
LG7	8924289	G/C	1,00	0,84	0,15	6458	(P<0,01)
LG7	8924291	C/-	1,00	0,84	0,15	6458	(P<0,01)
LG7	11248005	G/A	0,79	0,97	0,18	5830	(P<0,01)
LG7	14523935	A/G	0,66	1,00	0,34	3855	(P<0,01)
LG7	14523937	G/A	0,67	1,00	0,33	3855	(P<0,01)
scaf_195	200474	G/-	1,00	0,88	0,12	1268	(P<0,05)
scaf_195	200475	T/-	1,00	0,88	0,12	1268	(P<0,05)
scaf_195	200476	T/-	1,00	0,88	0,12	1268	(P<0,05)
scaf_235	69892	C/T	0,96	0,74	0,22	5729	(P<0,05)
scaf_266	111335	T/A	0,85	0,99	0,14	3111	(P<0,05)
scaf_350	136689	T/C	0,23	0,48	0,25	5023	(P<0,05)
scaf_350	136703	-/A	0,88	0,53	0,35	4704	(P<0,01)
scaf_350	136724	G/-	0,88	0,53	0,35	4703	(P<0,01)
scaf_766	7250	A/G	0,42	1,00	0,58	2196	(P<0,01)
scaf_766	7258	T/C	0,42	1,00	0,58	2195	(P<0,01)
scaf_959	36242	G/A	0,52	0,83	0,31	1189	(P<0,05)

scaf_959	36258	T/C	0,53	0,88	0,35	1189	(P<0,05)

Tabell 3 Bialleliske SNP funnet med Apek1 med signifikant ulik allelfrekvens mellom gruppene karakteriser for tidlig stengelstrekning og sen stengelstrekning.(P<0,05). Kromosomnummer Major / Minor er bestemt ut fra til hvilket allel som hadde flest reads i .vcf råfilen.

			Allelfrekvens Major				
Kromosom	Posisjon på kromosom	SNP Major/Minor	Tidlig	Sen	Differanse	Dybde	Signifikans
LG1	1582161	T / G	0,72	0,35	0,37	1383	(P<0,05)
LG1	9168543	C / A	0,96	0,82	0,14	713	(P<0,05)
LG1	9168576	T / C	0,96	0,82	0,14	713	(P<0,05)
LG1	10672991	T / A	0,79	1,00	0,21	1229	(P<0,05)
LG1	17230321	C / T	0,81	0,97	0,16	1594	(P<0,05)
LG1	21077660	T / A	0,86	0,99	0,13	3662	(P<0,05)
LG1	21918035	G / T	0,56	0,82	0,26	935	(P<0,05)
LG1	23217319	A / C	0,95	0,81	0,14	1178	(P<0,05)
LG2	8210726	G / C	0,76	0,44	0,32	796	(P<0,05)
LG2	13860187	A / G	1,00	0,76	0,24	1536	(P<0,05)
LG2	20407054	T / G	0,74	0,93	0,19	1441	(P<0,05)
LG2	20407055	A / C	0,74	0,93	0,19	1444	(P<0,05)
LG2	24137546	G / A	0,92	0,72	0,21	1185	(P<0,05)
LG3	2673758	G / A	0,63	0,91	0,28	928	(P<0,01)
LG3	2673760	T / C	0,89	0,36	0,53	928	(P<0,01)
LG3	3449274	C / T	1,00	0,88	0,12	1577	(P<0,05)
LG3	5423967	T / C	0,30	0,85	0,55	1129	(P<0,01)
LG3	16963320	A / C	0,99	0,87	0,12	2512	(P<0,05)
LG3	17682017	C / G	0,83	0,97	0,14	1310	(P<0,05)
LG3	17682035	A / G	0,83	0,97	0,14	1310	(P<0,05)
LG3	17682046	G / C	0,83	0,97	0,14	1310	(P<0,05)
LG4	1019735	G / A	0,99	0,87	0,11	5395	(P<0,05)
LG4	2858453	C / T	0,91	0,72	0,19	9374	(P<0,05)
LG4	3006300	C / T	0,83	0,97	0,14	3407	(P<0,05)
LG4	4840918	C / T	0,75	0,94	0,20	1982	(P<0,05)
LG4	4840919	T / G	0,75	0,94	0,20	1982	(P<0,05)
LG4	4850140	C / G	0,93	0,77	0,16	7583	(P<0,05)
LG4	5053247	G / A	0,79	0,49	0,30	767	(P<0,01)
LG4	22293793	A / T	0,73	0,49	0,24	962	(P<0,05)
LG4	22293794	T / C	0,73	0,49	0,25	962	(P<0,05)
LG5	906350	T / C	0,85	0,51	0,34	1368	(P<0,01)
LG5	1283650	C / A	0,78	0,97	0,20	757	(P<0,01)
LG5	1283651	A / C	0,78	0,97	0,20	757	(P<0,01)
LG5	1283653	A / T	0,78	0,97	0,20	757	(P<0,01)
LG5	3693778	T / -	0,91	0,70	0,21	980	(P<0,05)

LG5	3836559	C	/	-	0,99	0,89	0,11	996	(P<0,05)
LG5	3836560	A	/	-	0,99	0,89	0,11	996	(P<0,05)
LG5	3836561	T	/	-	0,99	0,89	0,11	996	(P<0,05)
LG5	3836562	A	/	-	0,99	0,89	0,11	996	(P<0,05)
LG5	6525196	C	/	A	0,79	0,97	0,18	752	(P<0,05)
LG5	7766776	A	/	G	0,82	0,95	0,13	1912	(P<0,05)
LG5	12053561	T	/	A	0,53	0,81	0,28	1311	(P<0,05)
LG5	12219661	T	/	-	0,55	0,74	0,19	3151	(P<0,05)
LG5	13001581	A	/	-	0,89	0,69	0,20	956	(P<0,05)
LG5	13001582	C	/	-	0,89	0,69	0,20	956	(P<0,05)
LG6	4049420	A	/	C	0,78	0,55	0,23	890	(P<0,05)
LG6	4049423	-	/	T	0,88	0,63	0,25	890	(P<0,05)
LG6	4207656	C	/	T	0,96	0,60	0,36	900	(P<0,01)
LG6	6113654	G	/	A	0,98	0,81	0,17	2639	(P<0,05)
LG6	7551402	T	/	C	0,95	0,80	0,14	3981	(P<0,05)
LG6	9427467	A	/	G	0,88	0,99	0,11	6879	(P<0,05)
LG6	10343986	C	/	G	0,77	0,95	0,17	4001	(P<0,01)
LG6	11814745	C	/	A	0,46	0,71	0,25	1011	(P<0,05)
LG6	12647432	C	/	T	0,80	0,98	0,18	3204	(P<0,01)
LG6	12647447	C	/	T	0,80	0,98	0,18	3204	(P<0,01)
LG6	15026256	A	/	G	1,00	0,89	0,11	1023	(P<0,05)
LG6	15280022	-	/	G	0,79	0,55	0,24	4471	(P<0,05)
LG6	15280041	A	/	-	0,79	0,56	0,23	4141	(P<0,05)
LG6	15280043	A	/	-	0,79	0,55	0,24	4471	(P<0,05)
LG6	15280068	-	/	G	0,79	0,56	0,23	4141	(P<0,05)
LG6	15723224	A	/	G	1,00	0,75	0,25	2703	(P<0,01)
LG6	16213525	C	/	G	0,36	0,66	0,30	2039	(P<0,01)
LG6	17939377	T	/	C	0,82	0,96	0,13	5018	(P<0,05)
LG6	18569896	T	/	C	0,93	0,76	0,18	918	(P<0,05)
LG6	19607788	G	/	A	0,98	0,83	0,15	2364	(P<0,05)
LG6	19951646	A	/	G	0,72	0,48	0,24	718	(P<0,05)
LG6	19951660	T	/	C	0,72	0,48	0,24	718	(P<0,05)
LG6	19951661	C	/	G	0,72	0,48	0,24	718	(P<0,05)
LG6	20401679	G	/	C	0,89	0,61	0,28	780	(P<0,01)
LG6	20401706	A	/	C	0,89	0,61	0,28	780	(P<0,01)
LG6	20871716	G	/	A	0,98	0,77	0,21	1349	(P<0,01)
LG6	20970059	T	/	C	0,97	0,82	0,15	776	(P<0,05)
LG6	21092252	G	/	A	0,57	0,80	0,23	2753	(P<0,05)
LG6	21093073	T	/	A	0,34	0,67	0,33	1680	(P<0,01)
LG6	21093080	C	/	A	0,50	0,76	0,27	1683	(P<0,05)
LG6	21093208	G	/	A	0,96	0,72	0,23	1778	(P<0,01)
LG6	21221809	G	/	T	0,98	0,83	0,15	2405	(P<0,05)
LG6	21881050	T	/	A	0,86	0,40	0,46	1532	(P<0,01)
LG6	22330396	G	/	T	1,00	0,89	0,11	1051	(P<0,05)

LG6	22529694	C	/	G	1,00	0,88	0,12	4608	(P<0,01)
LG6	22638089	T	/	G	0,59	0,82	0,24	2528	(P<0,05)
LG7	4688038	A	/	C	0,73	0,47	0,26	1751	(P<0,05)
LG7	4808860	A	/	G	0,49	0,71	0,21	1868	(P<0,05)
LG7	4877232	A	/	C	0,48	0,75	0,27	1396	(P<0,05)
LG7	6044309	C	/	T	0,97	0,79	0,19	1019	(P<0,05)
LG7	6063939	T	/	C	0,86	1,00	0,14	2878	(P<0,05)
LG7	6561357	T	/	G	1,00	0,89	0,11	1380	(P<0,05)
LG7	6603333	A	/	-	0,74	0,31	0,43	1380	(P<0,01)
LG7	6603349	T	/	C	0,33	0,77	0,44	2689	(P<0,01)
LG7	6603360	A	/	G	0,31	0,77	0,46	1380	(P<0,01)
LG7	8706752	A	/	-	0,73	0,95	0,22	973	(P<0,01)
LG7	8706753	A	/	-	0,73	0,95	0,22	973	(P<0,01)
LG7	8716678	A	/	G	0,89	0,98	0,09	5533	(P<0,05)
LG7	9925585	A	/	G	0,65	0,86	0,22	2204	(P<0,05)
LG7	10352653	G	/	A	0,59	0,34	0,25	4813	(P<0,05)
LG7	10352691	T	/	A	0,60	0,35	0,25	9333	(P<0,05)
LG7	15175628	A	/	T	1,00	0,78	0,21	4672	(P<0,01)
LG7	15789142	G	/	-	0,92	0,78	0,14	1290	(P<0,05)
LG7	15789144	A	/	-	0,92	0,78	0,14	1290	(P<0,05)
LG7	15789145	A	/	-	0,92	0,78	0,14	1290	(P<0,05)
LG7	18693372	G	/	T	0,98	0,86	0,13	2277	(P<0,05)
LG7	25085103	A	/	C	0,61	0,82	0,22	1051	(P<0,05)
LG7	26079080	G	/	C	0,25	0,48	0,23	12326	(P<0,05)
LG7	26079113	T	/	C	0,75	0,52	0,23	12321	(P<0,05)
LG7	28474812	G	/	T	0,65	0,86	0,21	6207	(P<0,05)
LG7	28641323	T	/	C	1,00	0,91	0,09	1643	(P<0,05)
LG7	28847783	-	/	T	1,00	0,92	0,08	2566	(P<0,05)
scaf_74	569894	G	/	A	1,00	0,85	0,15	1178	(P<0,05)
scaf_82	143980	G	/	C	0,57	0,82	0,25	2745	(P<0,05)
scaf_82	515442	C	/	T	0,86	0,39	0,48	3862	(P<0,01)
scaf_119	13174	C	/	T	0,82	0,34	0,49	822	(P<0,01)
scaf_119	255853	A	/	C	0,94	0,75	0,19	6179	(P<0,05)
scaf_152	121219	C	/	A	0,54	0,82	0,28	8216	(P<0,01)
scaf_164	313499	T	/	C	0,97	0,83	0,14	2734	(P<0,05)
scaf_164	313502	T	/	A	0,97	0,83	0,14	2734	(P<0,05)
scaf_174	148835	T	/	C	0,47	0,77	0,29	2977	(P<0,05)
scaf_186	121970	T	/	C	0,93	0,79	0,13	7715	(P<0,05)
scaf_193	235740	T	/	G	0,63	0,90	0,27	1307	(P<0,05)
scaf_212	147227	A	/	-	0,69	0,48	0,21	1550	(P<0,05)
scaf_212	147260	-	/	A	0,69	0,48	0,21	1550	(P<0,05)
scaf_215	174301	C	/	T	0,93	0,75	0,18	2480	(P<0,05)
scaf_255	263975	G	/	A	0,52	0,79	0,27	2935	(P<0,05)
scaf_267	169958	G	/	A	0,53	0,33	0,21	1434	(P<0,05)

scaf_267	267337	G	/	A	0,76	1,00	0,24	1323	(P<0,01)
scaf_320	70770	T	/	A	0,85	1,00	0,15	963	(P<0,01)
scaf_407	47368	C	/	T	0,77	0,53	0,25	1446	(P<0,05)
scaf_442	72281	A	/	C	0,85	1,00	0,15	1635	(P<0,01)
scaf_457	110876	C	/	T	0,91	0,76	0,14	2249	(P<0,05)
scaf_512	53515	C	/	T	0,78	0,94	0,16	1602	(P<0,05)
scaf_512	170959	C	/	T	1,00	0,83	0,17	865	(P<0,01)
scaf_590	103114	T	/	G	0,18	0,41	0,22	7288	(P<0,05)
scaf_668	105068	G	/	A	0,60	0,80	0,20	2784	(P<0,05)
scaf_670	82529	G	/	A	0,70	0,42	0,28	1110	(P<0,05)
scaf_700	31116	T	/	G	0,86	1,00	0,14	2606	(P<0,01)
scaf_774	21891	C	/	T	0,84	0,98	0,15	904	(P<0,05)
scaf_774	62738	A	/	-	0,79	0,94	0,16	1986	(P<0,05)
scaf_774	62787	-	/	G	0,79	0,94	0,16	1986	(P<0,05)
scaf_806	88642	T	/	G	1,00	0,87	0,13	2624	(P<0,05)
scaf_885	9136	G	/	A	0,99	0,87	0,12	2102	(P<0,05)
scaf_893	28441	A	/	G	0,84	1,00	0,16	777	(P<0,01)
scaf_956	7118	A	/	T	0,62	0,86	0,24	2505	(P<0,01)
scaf_956	7119	T	/	G	0,62	0,86	0,24	2505	(P<0,01)
scaf_962	29984	C	/	T	0,79	0,95	0,16	2228	(P<0,05)
scaf_977	45297	A	/	G	0,95	0,79	0,16	1579	(P<0,05)
scaf_977	45340	A	/	-	0,91	0,76	0,15	1579	(P<0,05)
scaf_2688	2677	-	/	A	1,00	0,90	0,09	4707	(P<0,05)
scaf_2688	2708	A	/	-	1,00	0,90	0,10	4707	(P<0,05)
scaf_2722	4830	A	/	T	0,80	0,96	0,15	1414	(P<0,05)
scaf_2856	2938	G	/	A	1,00	0,81	0,19	2445	(P<0,01)
scaf_3591	2926	A	/	T	0,78	1,00	0,22	1238	(P<0,05)
scaf_4084	1276	G	/	-	0,92	0,74	0,18	1754	(P<0,05)
scaf_4084	1277	A	/	-	0,92	0,74	0,18	1754	(P<0,05)
scaf_5682	141	G	/	A	0,90	1,00	0,10	7186	(P<0,05)
scaf_6992	1002	G	/	A	0,47	0,76	0,28	1329	(P<0,01)
scaf_8648	234	C	/	T	0,99	0,87	0,12	2362	(P<0,05)
scaf_14914	946	-	/	C	0,93	0,77	0,15	3294	(P<0,05)
scaf_24204	594	G	/	A	0,86	0,99	0,13	1622	(P<0,05)
scaf_24204	595	T	/	A	0,86	0,99	0,13	1622	(P<0,05)
scaf_28709	135	C	/	A	0,97	0,83	0,14	1131	(P<0,01)
scaf_36286	233	C	/	T	0,64	0,91	0,27	1384	(P<0,01)

Tabell 4 Trialleliske SNP funnet med signifikant (P<0,05) ulik allelfrekvens mellom gruppene karakterisert for tidlig og sent tidspunkt for stengelstrekning. Formelen Fst er egentlig laget for å regne frekvens mellom 2 alleler. Fst retning angir hvilke allelene frekvenser: Major, minor1 eller minor2, som er slått sammen i par og sammenlignet med den tredje. Posisjon på kromosomet er anngitt til venstre. Neste kolonne viser major og minor frekvens av SNP ut fra hva som hadde mest reads totalt(Major). Gjennomsnittlig allelfrekvens innen gruppene og differansen mellom snittene er vist under differanse

APEK1				**Allelfrekvens i del1 av fst**				
Fst Retning	**Posisjon**	**Major**	**M1,M2**	**SENE**	**TIDLIGE**	**diff**	**Signifikans**	**Dybde**
Ma ≠ (Mi1+Mi2)	LG1-14785961	G	C,T	0,61	0,38	0,22	(P<0,05)	3903
(Ma+Mi1) ≠ Mi2	LG1-23217318	C	A,T	0,95	0,80	0,14	(P<0,05)	1178
(Ma+Mi1) ≠ Mi2	LG1-27863009	T	-,A	1,00	0,92	0,08	(P<0,05)	1339
(Ma+ Mi2) ≠ Mi1	LG2-515715	G	A,T	0,69	0,50	0,19	(P<0,05)	763
(Ma+Mi1) ≠ Mi2	LG2-20407053	T	A,C	0,74	0,93	0,19	(P<0,05)	1444
(Ma+ Mi2) ≠ Mi1	LG4-917816	G	C,T	0,98	0,82	0,16	(P<0,05)	1659
(Ma+ Mi2) ≠ Mi1	LG4-1019565	A	C,T	0,79	0,59	0,20	(P<0,05)	5082
(Ma+Mi1) ≠ Mi2	LG4-7791668	G	A,C	1,00	0,92	0,08	(P<0,05)	1104
Ma ≠ (Mi1+Mi2)	LG5-12219691	-	T,G	0,55	0,74	0,19	(P<0,05)	6302
(Ma+Mi1) ≠ Mi2	LG6-18406170	T	-,C	0,94	1,00	0,06	(P<0,05)	723
Ma ≠ (Mi1+Mi2)	LG6-18890904	A	T,G	0,88	0,68	0,20	(P<0,05)	1787
(Ma+Mi1) ≠ Mi2	LG7-14066942	C	G,T	0,91	1,00	0,09	(P<0,05)	2027
Ma ≠ (Mi1+Mi2)	LG7-15789143	T	-,C	0,92	0,77	0,15	(P<0,05)	1290
(Ma+ Mi2) ≠ Mi1	LG7-15789143	T	-,C	0,92	0,78	0,14	(P<0,05)	1290
Ma ≠ (Mi1+Mi2)	LG7-17998574	G	T,A	0,66	0,36	0,30	(P<0,01)	1973
(Ma+ Mi2) ≠ Mi1	LG7-17998574	G	T,A	0,83	0,51	0,32	(P<0,01)	1973
Ma ≠ (Mi1+Mi2)	scaf_702-112119	G	-,T	1,00	0,88	0,12	(P<0,01)	1613
Ma ≠ (Mi1+Mi2)	scaf_912-13065	T	G,A	0,84	0,59	0,25	(P<0,05)	1074
Ma ≠ (Mi1+Mi2)	scaf_1017-42268	T	G,C	1,00	0,91	0,09	(P<0,05)	1492
Ma ≠ (Mi1+Mi2)	scaf_2623-2286	G	C,A	0,55	0,91	0,35	(P<0,01)	1602
(Ma+Mi1) ≠ Mi2	scaf_16044-479	C	G,T	0,76	0,98	0,22	(P<0,01)	1108

Pst1								
Fst Retning	**Posisjon**	**Major**	**M1,M2**	**SENE**	**TIDLIGE**	**Diff**	**Signifikans**	**Dybde**
(Ma+ Mi2) ≠ Mi1	LG2-19944574	T	A,4	0,65	0,92	0,26	(P<0,05)	5132
Ma ≠ (Mi1+Mi2)	LG3-30484546	-	T,A	0,64	0,98	0,34	(P<0,01)	5848
(Ma+ Mi2) ≠ Mi1	LG3-30484546	-	T,A	0,64	0,99	0,35	(P<0,01)	5848
(Ma+Mi1) ≠ Mi2	LG6-21393294	-	T,C	0,86	0,45	0,41	(P<0,01)	7951
(Ma+ Mi2) ≠ Mi1	LG6-21393294	-	T,C	0,47	0,85	0,38	(P<0,01)	7951

Vedlegg 5

Pearl-script for filtrering

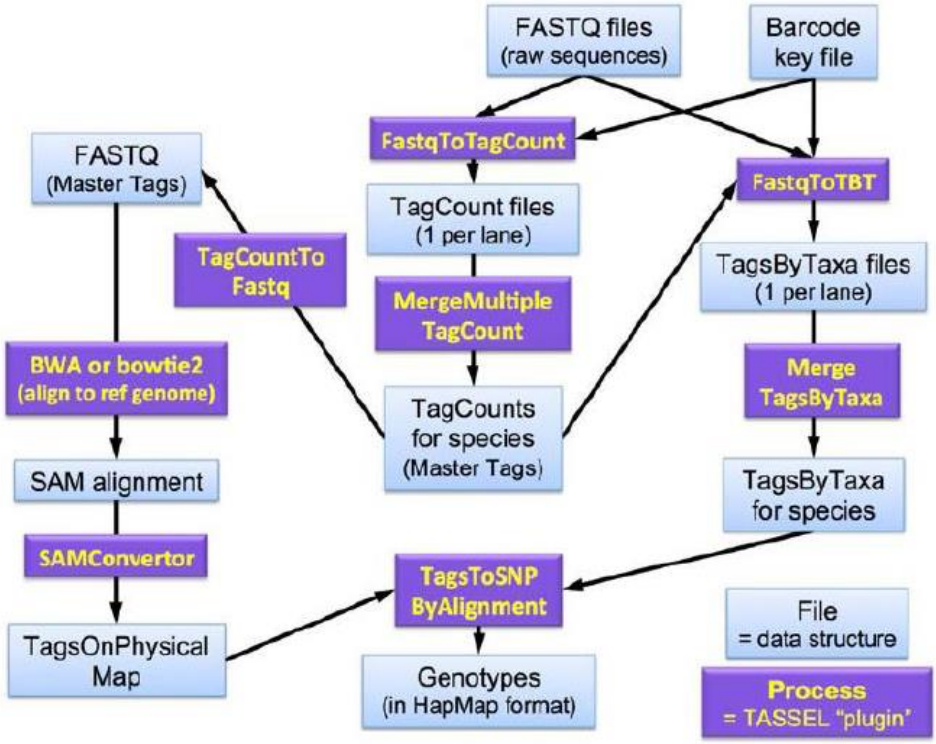
Sletter begge alleler for prøve dersom en av de har 127 reads, krav på minimum 100 reads over hele pool og Minor allele frequency >0,05 for hele SNP. Skriptet under hører til Apek1. En liten modifikasjon måtte gjøres for bruk i Pst1 på linje 20 som angir størrelse på poolene.

```
#/bin/perl

$start = 0;
$count = 0;
while(<>) {
   chop;
   @arr = split(/\t/);
   if ($arr[0] eq "#CHROM") {
     @head = @arr;
     print $_, "\n";
     $start = 1;
     for (@head) { $count++; };
   } elsif ($start > 0) {
     $total = 0;
     $tsumref = 0;
     $tsumalt = 0;
                @groups = (0,0,0,0,0,0);
                @sumref = (0,0,0,0,0,0);
                @sumalt = (0,0,0,0,0,0);
                @sizes = (16,16,16,15,16,16);

     $pos = 9;
     for ($group =0; $group <6; $group++) {
       for ($siz =0; $siz < $sizes[$group]; $siz++) {
         @elem = split(/[:]/, $arr[$pos++]);
         $num = 0; for (@elem) { $num++; };
         if ($num > 0) {
#          print $elem[1], "\n";
           @selem = split(/[,]/, $elem[1]);
           $snum = 0; for (@selem) { $snum++; };
           $rb = $selem[0];
           $ab = $selem[1];
           $xb = 0;
           if ($snum == 3) { $xb = $selem[2]; }
           # I linjen under filtrerer vi bort alle som er 127   && betyr AND,  || betyr OR
           if ($rb<127 && $ab <127 && $rb + $ab > -1) {
               $ab = $ab + $xb;
               $sumref[$group] += $rb;
               $sumalt[$group] += $ab;
               $tsumref += $rb;
               $tsumalt += $ab;
               $groups[$group]++;
               $total++;
           }
         }
       }
     }

     $lim = 10;
     $req1 = $total > 0 && $tsumref > 0 && $tsumalt > 0 && $tsumref / ($tsumref + $tsumalt) >= 0.05 && $tsumalt / ($tsumref + $tsumalt) >= 0.05;
     # her er hver av gruppene skal være  minst 5% av totalen (må først sjekke at tallene er > 0)
     $req2 = $sumref[0] + $sumalt[0] >= 100 &&
             $sumref[1] + $sumalt[1] >= 100 &&
             $sumref[2] + $sumalt[2] >= 100 &&
             $sumref[3] + $sumalt[3] >= 100 &&
             $sumref[4] + $sumalt[4] >= 100 &&
```

```
57              $sumref[5] + $sumalt[5] >= 100; # hver av guppene skal være minst 100
58        if ($req1 && $req2)  {
59          # På linjene under kan du prøve ulike format.
60          #  # betyr kommentar,
61          # Linje 1 printer første 9 kolonner
62          # Linje 2 printer hvor mange av de to gruppene som har tall under 127 i begge grupper
63          # Linje 3 printer alle kolonner etter kolonne 9
64          # Linje 4 printer sumref / sumalt for hver av de 6 gruppene
65          # linje 5 printer <linjeskift>
66          # ring om du lurer på noe
67          for ($e=0; $e<9; $e++) { print $arr[$e], "\t"; };
68          #printf ("%d,%d %d,%d %d,%d %d,%d %d,%d %d,%d\n", $groups[0], $groups[1], $groups[2], $groups[3], $groups[4], $groups[5]);
69          # for ($e=9; $e<$count; $e++) { print $arr[$e], "\t"; };
70          for ($e=9; $e<$count; $e++) {
71           @elem = split(/[:]/, $arr[$e]);
72           $num = 0; for (@elem) { $num++; };
73           $rb = $ab = $xb = "";
74           if ($num > 0) {
75             @selem = split(/[,]/, $elem[1]);
76             $snum = 0; for (@selem) { $snum++; };
77              $rb = $selem[0];
78              $ab = $selem[1];
79              if ($snum == 3) {
80                $xb = $selem[2];
81              }
82              if ($rb == 127 || $ab ==127 ||  $xb ==127) { $rb = $ab = $xb = ""; };
83           }
84            print $rb, "\t", $ab, "\t", $xb, "\t";
85
86          }
87          #printf ("%d,%d %d,%d %d,%d %d,%d %d,%d %d,%d", $sumref[0], $sumalt[0], $sumref[1], $sumalt[1], $sumref[2], $sumalt[2], $sumref[3], $sumalt[3],$sumref[4], $sumalt[4],$sumref[5], $sumalt[5]);
88        print "\n";
89        }
90      } else {
91        print $_, "\n";
92      }
93
94  }
95
```

Vedlegg 6

Utdrag fra GBS-rapportene fra Genomic Diversity Facility, Cornell University



Figur 1 Flytskjemaet over viser hvordan stegene i en mulig GBS «Discovery Pipeline» analyse henger sammen (variasjoner til denne tilnærmingen er mulig). Lyseblå bokser representerer filer (eller datastrukturer) produsert på hvert steg i analysen og de lilla boksene representerer prosessene (Tassel 3 plugins) som produserer dem.

Apek1

Tabell 1 Apek1, sammendrag over Reads og Tags funnet i hver sekvenserings-«lane»

File	Barcodes	Reads	Good_Barcoded_Reads	Tags
H5GTLBGXY_1_fastq.gz	25	664849166	78233780	5314452
C9CRUANXX_1_fastq.gz	71	267566302	248553328	17897572

Tabell 2 Apek1, Resultat av alignment mot referansegenomet utført med bwa

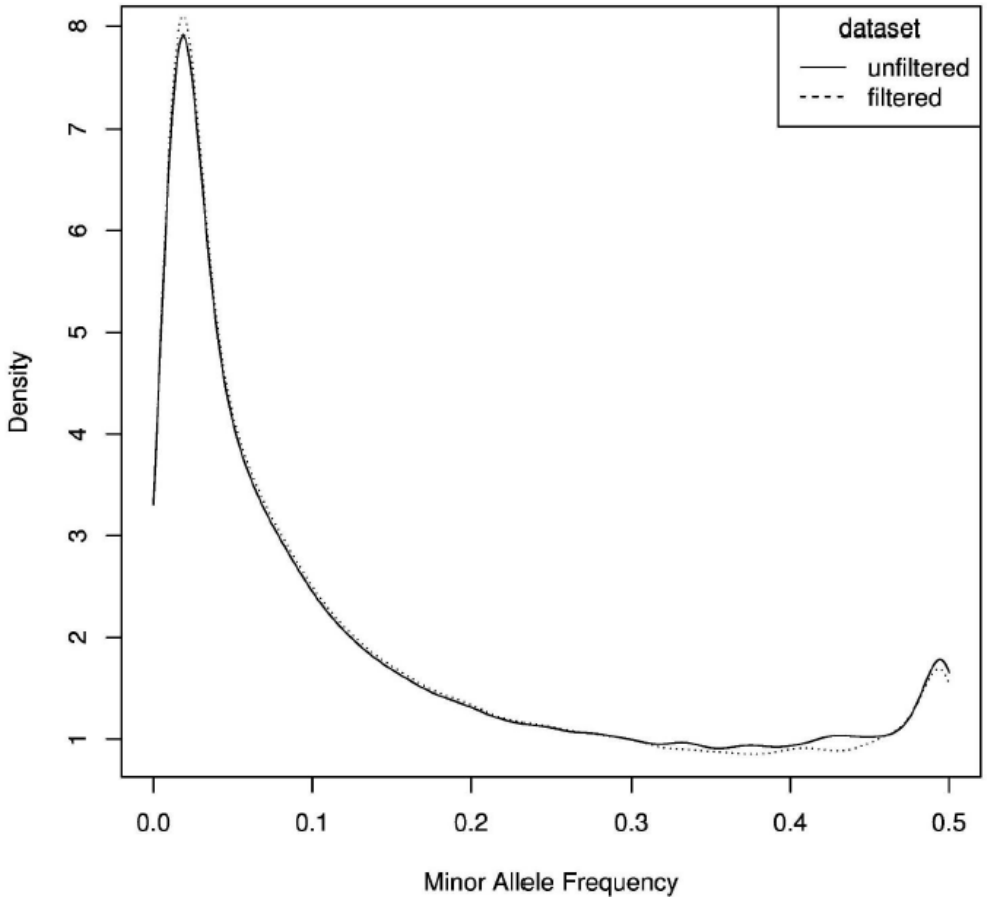
category	number	percent
total	2788388	100
uniquely_aligned	1657584	59.4
multiply_aligned	99777	3.6
unaligned	1031027	37.0

Tabell 3 Apek1, Antall SNPer funnet

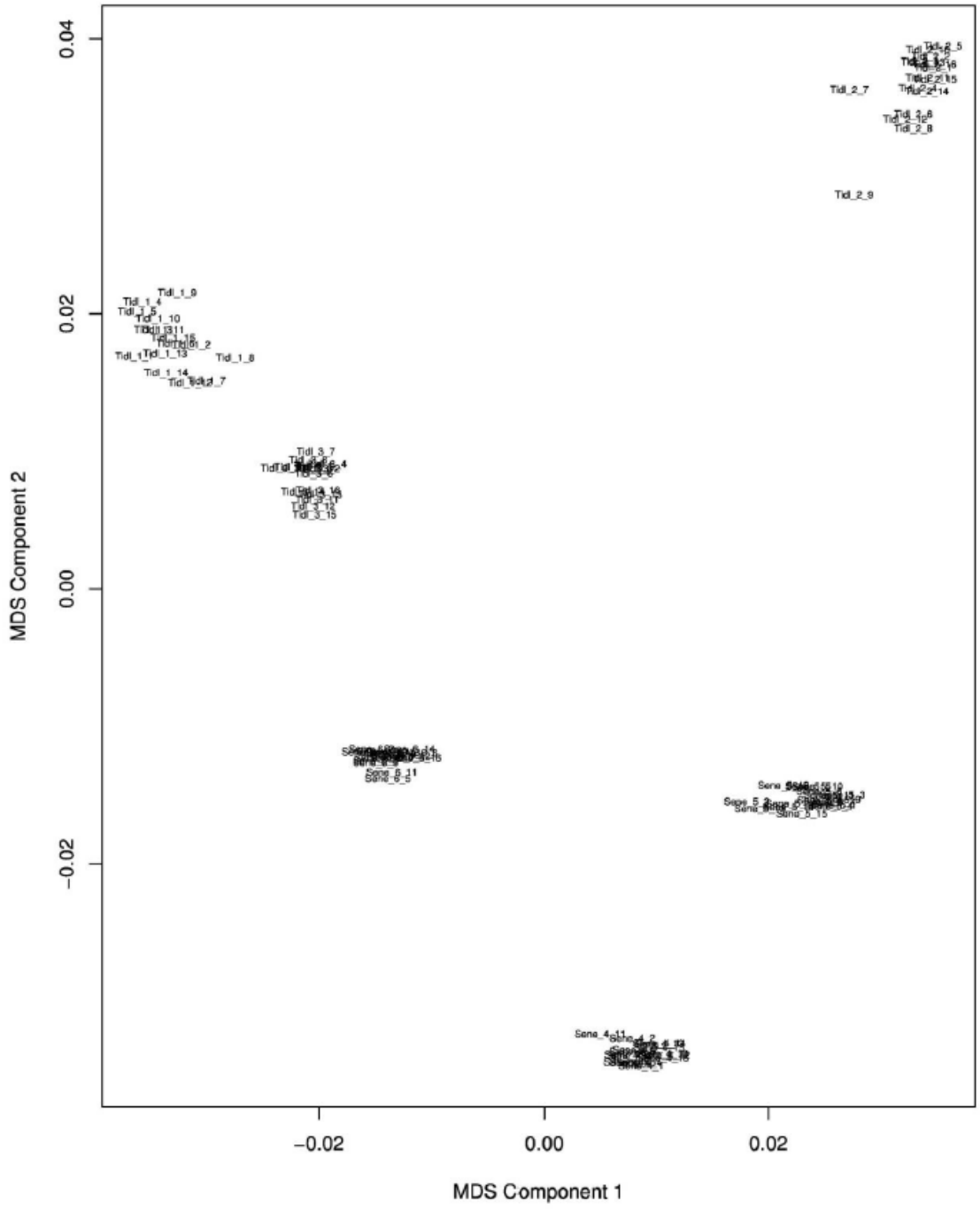
Råfil (all.vcf)	284967
Filtrert (all.filtered.recode.vcf)	276576

Tabell 4 Apek1, Dybde og missingness for .vcf filene. Ufiltrert: all.vcf og filtrert (MAF>0,01 og Missing sites <90%)

All.vcf.gz			
	mean	median	standard_deviation
individual depth	17.214	17.954	4.154
site depth	14.384	5.716	24.189
individual missingness	0.212	0.192	0.096
site missingness	0.212	0.042	0.282
all.filtered.recode.vcf.gz			
	mean	median	standard deviation
individual depth	17.245	17.843	3.801
site depth	14.458	5.809	24.214
individual missingness	0.192	0.179	0.051
site missingness	0.192	0.032	0.271



Figur 2 Apek1, Minor allelfrekvens distribusjon for bialleliske loci



Figur 3 Apek1, Multi dimensional scaling plot (MDS) over bialleliske SNPer. Tidlige og sene pooler er adskilt. Tidlige pooler ligger på den positive delen av «MDS component 2» aksen mens sene poler ligger på den negative delen av «MDS component 2» aksen.

Pst1

Tabell 5 Pst1, sammendrag over Reads og Tags funnet i hver sekvenserings-«lane»

File	Barcodes	Reads	Good_Barcoded_Reads	Tags
C9CB5ANXX_3_fastq.gz	96	263146346	229335515	8246039

Tabell 6 Pst1, Resultat av alignment mot referansegenomet utført med bwa

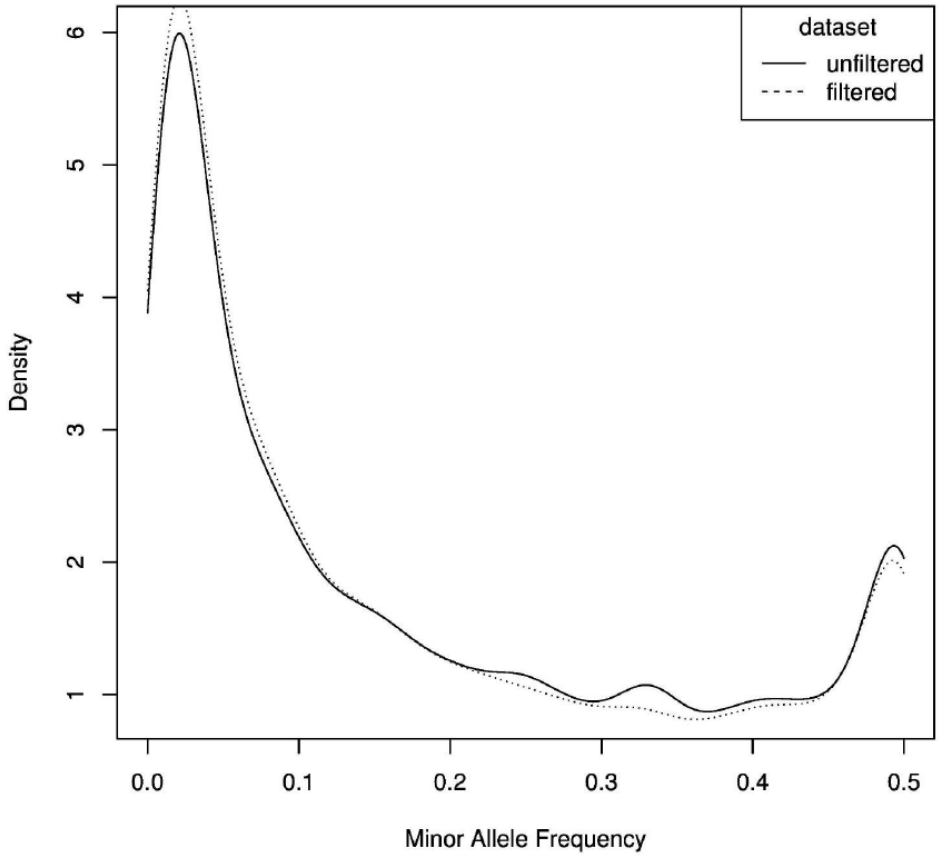
category	number	percent
total	1237695	100
uniquely_aligned	809982	65.4
multiply_aligned	43393	3.5
unaligned	384320	31.1

Tabell 7 Pst1, Antall SNPer funnet

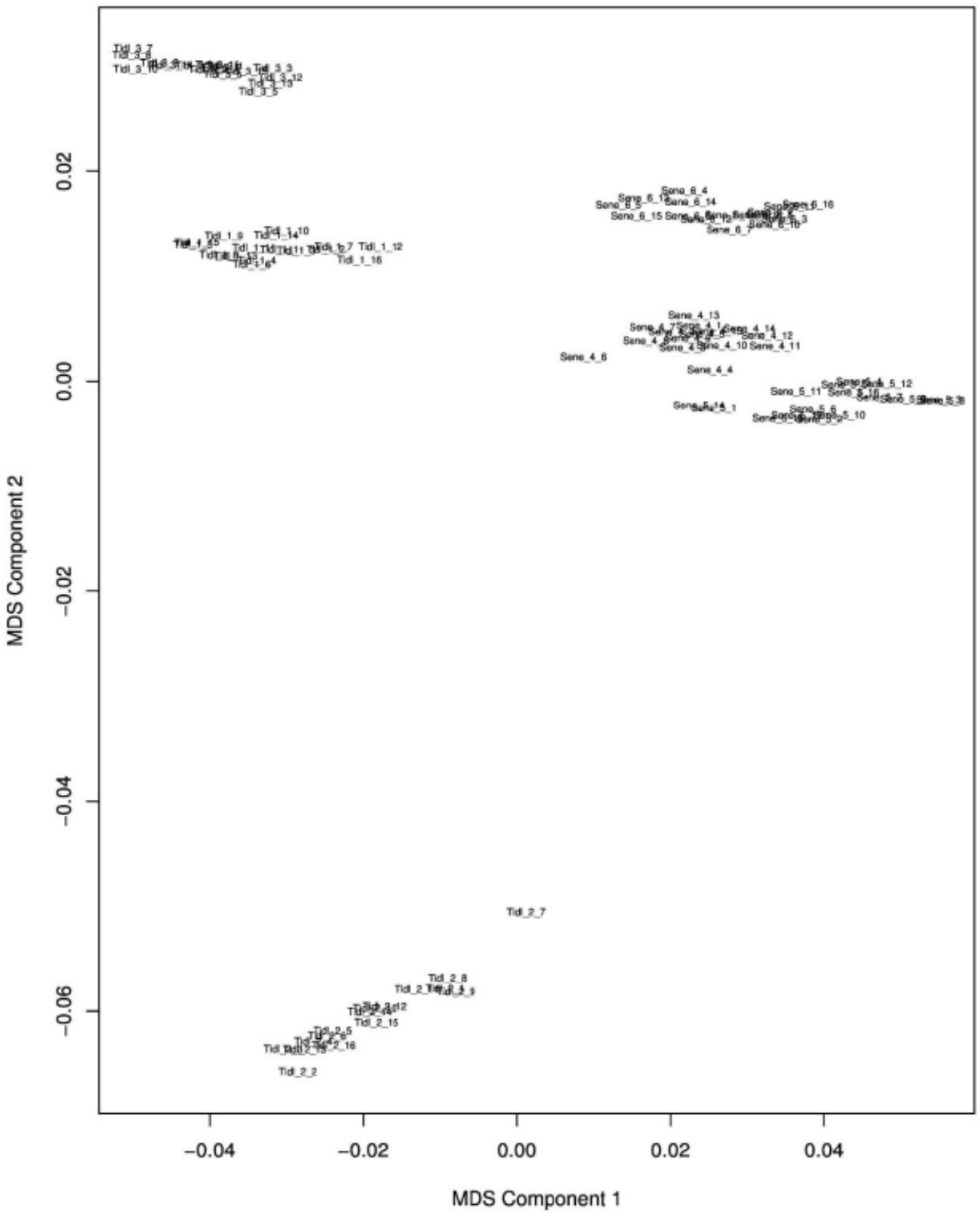
Råfil (all.vcf)	34713
Filtrert (all.filtered.recode.vcf)	32553

Tabell 8 Pst1, Dybde og missingness for .vcf filene. Ufiltrert: all.vcf og filtrert (MAF>0,01 og Missing sites <90%)

All.vcf.gz			
	mean	median	standard_deviation
individual depth	56.815	57.440	9.268
site depth	47.992	12.358	66.126
individual missingness	0.191	0.175	0.097
site missingness	0.191	0.031	0.284
all.filtered.recode.vcf.gz			
	mean	median	standard_deviation
individual depth	57.600	57.743	7.412
site depth	49.974	13.234	67.105
individual missingness	0.151	0.142	0.054



Figur 4 Pst1, Minor allelfrekvens distribusjon for bialleliske loci



Figur 5 Pst1, Multi dimensional scaling plot (MDS) over bialleliske SNPer. Tidlige og sene pooler er adskilt. Tidlige pooler ligger på den negative delen av «MDS component 1» aksen, mens sene poler ligger på den positive delen av «MDS component 1» aksen.

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

Postboks 5003
NO-1432 Ås
Norway