NTNU
Kunnskap for en bedre verden

Faglig sluttrapport

FHF-prosjektnummer: 901340

Prosjekttittel: **Redusert ferskvannsoppgang hos oppdrettslaks?**

Dato: 15.11.19

Utfylt av: Stig W. Omholt, Laila Berg, Kjetil Hindar, Geir H. Bolstad og Sigbjørn Lien, med bidrag fra hele prosjektgruppen

Sammendrag

Oppgangen av oppdrettslaks i norske elver forårsaker negative konsekvenser for villakspopulasjonene og negativ omtale av norsk oppdrettsnæring. I dette prosjektet har vi undersøkt om det finnes genetiske holdepunkter som sannsynliggjør at det kan være mulig å avle frem laks som i liten eller ingen grad vil søke opp i elvene for å gyte dersom de rømmer. Prosjektet er et samarbeid mellom NTNU sitt Senter for biodiversitetsdynamikk, Senter for Integrert Genetikk (CIGENE) ved NMBU, og NINA sine lakseforskere i Trondheim. Omlag 6000 rømte oppdrettslaks vandrer hvert år i fiskesesongen opp i norske lakseelver fra en årlig estimert totalpopulasjon på mellom 15 000 - 910 000 individer. En mulig fortolkning av denne kontrasten er at den store andelen rømt oppdrettslaks som ikke vender tilbake til ferskvann iallfall delvis kan skyldes 12-15 generasjoner med avlsarbeid. Målsettingen med prosjektet har derfor vært å undersøke om det er genetiske forskjeller mellom oppdrettspopulasjonen og oppdrettslaks fanget i norske lakseelver, om disse forskjellene beror på at oppdrettslaks fanget i elv er mer lik villaks enn annen oppdrettslaks, og om disse eventuelle forskjellene kan kobles til biologiske mekanismer som underligger laksens evne til å vende tilbake til ferskvann. To tusen rømte oppdrettslaks, fanget i 98 forskjellige lakseelver fordelt utover norskekysten og identifisert som rømt oppdrettslaks på bakgrunn av skjellkarakterer, ble sammenliknet med i alt nær 800 oppdrettslaks fra de fire største avlsselskapene i Norge: AquaGen, Mowi, SalmoBreed og Raumastammen (SalMar), og med omlag 1000 villaks fra 54 norske lakseelver. Sammenligningen ble gjort ved bruk av 48 000 mutasjoner (SNPer) fordelt over laksens kromosomer. Vi fant genetiske forskjeller mellom den generelle oppdrettspopulasjonen og oppdrettslaks fanget i norske lakseelver. Oppdrettslaks fanget i elv fra fire ulike avlspopulasjoner hadde et betydelig antall SNP-alleler felles, der de var mer lik villaks enn hva som kan tilskrives tilfeldigheter. Vi fant også at de genetiske forskjellene mellom rømt oppdrettslaks fanget i elv og avlspopulasjonene de kommer fra kan kobles til biologiske mekanismer som med stor sannsynlighet underligger laksens evne til å vandre opp i ferskvann. Resultatene indikerer at det kan være mulig å redusere den norske oppdrettspopulasjonens evne til å vandre opp i ferskvann ved bruk av molekylærgenetisk informert avlsarbeid. Vi konkluderer med at resultatene legitimerer en oppfølging som fjerner usikkerhetene forbundet med denne innledende studien.

Summary

Immigration of escaped farmed Atlantic salmon to Norwegian rivers has negative impact on wild salmon populations and harms the reputation of the fish-farming industry. In this project, we have investigated whether there is genetic evidence indicating that it might be possible to breed a salmon not capable of river entry. The project is a collaboration between NTNU Centre for Biodiversity Dynamics, NMBU Center for Integrative Genetics and the salmon research group at the Norwegian Institute for Nature Research (NINA). Only about 6000 escapees migrate annually into Norwegian rivers from an annual escapee population of 15 000 to 910 000 individuals. One possible interpretation of this contrast is that the very large proportion of escapees not showing up in fresh water is at least partially due to 12-15 generations of breeding work. The goal of this project has therefore been to investigate whether there are clear genetic differences between the farmed population and the escapee population, whether these possible differences resides in escapees being more similar to wild salmon than the farmed populations they come from, and whether these possible differences can be linked to biological mechanisms underlying the capacity for river entry. Two thousand escaped farmed salmon caught in 98 rivers along the Norwegian coast and identified as escaped farmed salmon from growth patterns in the scales, were compared with near 800 farmed salmon representing the four major fish breeding companies in Norway: AquaGen, Mowi, SalmoBreed and the Rauma strain (SalMar), and with about one thousand wild salmon from 54 Norwegian rivers. The comparisons were done by using 48,000 mutations (SNPs) distributed across the salmon genome. We found clear genetic differences between the general farmed salmon population (represented by samples from the major breeding lines) and escapees from these breeding lines caught in Norwegian salmon rivers. Farmed salmon caught in rivers from all four major breeding lines had a high number of common SNP alleles where they were more similar to wild salmon than can be explained by genetic drift alone. The genetic differences between escaped farmed salmon caught in rivers and the breeding lines they originated from can be connected to biological mechanisms which very probably influence the ability of salmon to enter fresh water to spawn. The results also indicate that it might be possible to reduce the number of escapees maintaining the capacity for river entry by genome-based precision breeding. We conclude that the results justify a more thorough investigation aimed at removing the uncertainties attached to this preliminary study.

1.0 Innledning

1.1 Bakgrunn

Oppgangen av oppdrettslaks i norske elver forårsaker negative konsekvenser for villakspopulasjonene og skaper negativ omtale for den norske oppdrettsnæringen. Selv om næringen bestreber seg på å utvikle rømningssikre produksjonsbetingelser er det vanskelig å sikre seg hundre prosent mot rømming med dagens produksjonsregime, og det er derfor behov for å komme opp med tiltak som reduserer problemet vesentlig selv om oppdrettslaks kommer fri i sjøen.

Omlag 6000 rømte oppdrettslaks vandrer hvert år i fiskesesongen opp i norske lakseelver fra en årlig estimert totalpopulasjon på mellom 15 000 - 910 000 individer (Fiskeridirektoratets rømmingsstatistikk, 2006 til 2018). Det er tre forklaringer på denne kontrasten. Den ene er at den skyldes summen av en hel rekke omstendigheter knyttet til rømningstidspunkt, sult, sykdommer og predasjon som ikke er årsaksmessig koblet til at oppdrettslaksen har vært avlet på i flere tiår. Den andre er at oppdrettslaksen gjennom 12-15 generasjoner med avlsarbeid har blitt selektert for egenskaper som indirekte forårsaker at majoriteten av den rømte oppdrettslaksen ikke vender tilbake til ferskvann. Den tredje er at kontrasten skyldes en kombinasjon av de to første forklaringene.

En testbar slutning fra det andre og tredje forklaringsalternativet er at de rømte oppdrettslaksene man finner igjen i ferskvann har en genotypisk signatur av betydning for ferskvannsoppvandring som har større likhet med de man finner i villaks enn de man finner i en stor andel av oppdrettslaks. Det vil si at disse individene har en genetisk sammensetning som ikke er representativ for de oppdrettsstammene de kommer fra, og derved i betydelig grad representerer en selektert populasjon.

Dette prosjektet har hatt som mål å etterprøve denne slutningen ved å gjennomføre en genetisk sammenligning av villaks, oppdrettslaks fra fire norske avlsstammer og oppdrettslaks fanget i norske elver, og gitt et utfall i samsvar med forventningen, undersøke om de avdekkede genetiske forskjellene er koblet til biologi av relevans for evnen til å vende tilbake til ferskvann. Et positivt utfall på denne etterprøvingen sannsynliggjør muligheten for at en ved hjelp av genombasert presisjonsavl vil kunne redusere oppvandringsevnen til rømt oppdrettslaks.

1.2 Prosjektets omfang

Hovedmålet med prosjektet har vært å avklare (i) om det finnes genetiske forskjeller mellom den generelle oppdrettspopulasjonen og oppdrettslaks som er fanget i norske lakseelver, (ii) om disse eventuelle forskjellene beror på at oppdrettslaks fanget i elv er mer lik villaks enn bakgrunnspopulasjonen av oppdrettslaks, og (iii) om de eventuelle genetiske forskjellene kan kobles til biologiske mekanismer man mener underligger laksens evne til å vende tilbake til ferskvann etter rømming.

Prosjektet har nyttiggjort seg data produsert av flere andre prosjekter gjennomført av NMBU og NINA finansiert av NFR, FHF og andre: (i) Genotype-data for 934 villaks, genotype-data fra 230 individer fra AquaGen's 2011 avlspopulasjon, og 50-80 individer fra avlspopulasjonene til Salmobreed, Mowi og SalMar (Rauma-stammen), alle genotypet med 220K SNP array utviklet for Atlantisk laks (*Salmo salar*). (ii) Helgenomsekvensdata for oppdrett og villaks. (iii) Algoritmer for å tilordne rømt oppdrettslaks til mest sannsynlig opphavspopulasjon. (iv) Data om SNPer som med stor sannsynlighet har vært under seleksjon i avlsstammen til AquaGen.

Prosjektet har bestått av følgende 8 delaktiviteter:

1. Gjennomgang av NINAs samling av skjell tatt fra oppdrettslaks fanget i elver langs hele norskekysten (>5000), utvelging av 2000 representative individer fra denne samlingen basert på innsamlingssted, skjellkvalitet og individenes forhistorie ut fra skjellanalyser.

2. Ekstraksjon av DNA fra de utvalgte 2000 skjellprøvene.

3. Utvikling av et nytt 60K array (SNP-chip) for storskala genotyping designet for dette prosjektet, samtidig som det er egnet for videre bruk av næringen så vel som til andre forskningsformål.

4. Genotyping av de 2000 utvalgte oppdrettslaksene fanget i elv med det nye SNP-arrayet.

5. Ekstraksjon av DNA og genotyping av 120 individer fra avlsstammene Mowi, SalmoBreed og Rauma for å få et sammenligningsgrunnlag for disse populasjonene på lik linje med allerede eksisterende data for AquaGen.

6. Tilordning av de 2000 utvalgte oppdrettslaksene fanget i elv til avlsstamme.

7. Analyse av genotype-data for å avklare om det finnes genetiske forskjeller mellom den generelle oppdrettspopulasjonen og oppdrettslaks som er fanget i norske lakseelver, og om disse eventuelle forskjellene beror på at oppdrettslaks fanget i elv er genetisk mer lik villaks enn oppdrettslaks.

8. Betinget av positivt utfall av aktivitet 7, belyse om påviste genetiske forskjeller kan kobles til biologiske mekanismer som påvirker laksens evne til å vende tilbake til ferskvann.

1.3 Prosjektorganisering

Stig W. Omholt, NTNU, har vært prosjektleder, og prosjektet har formelt vært forankret i Centre for Biodiversity Dynamics (CBD) ved NTNU (Senter for fremragende forskning). I tillegg til Omholt har prosjektgruppen inkludert Kjetil Hindar, NINA, og Sigbjørn Lien, CIGENE/NMBU. Disse tre har også utgjort styringsgruppen for prosjektet. Dr. Laila Berg, koordinator for NTNU Bioteknologi, har fungert som sekretær for styringsgruppen.

For dataanalysene har prosjektgruppen blitt utvidet til en analysegruppe med forskere fra NMBU (Nicola Barson) og NINA (Ingerid Julie Hagen Arnesen, Geir Bolstad og Sten Karlsson). Øvrige bidragsytere til prosjektet har vært: NMBU: Matthew Kent, Torfinn Nome og Silje Karoliussen; NINA: Gunnel Østborg, Sigrid Skoglund, Merethe Spets, Hege Brandsegg, Line Birkeland Eriksen, Sten Even Erlandsen og Bente Uhre Halvorsen; NTNU: Stig Omholt fikk hjelp av Oda Omholt (gjennom et treukers-engasjement) til å gjøre programvarekoden han utviklet for å analysere dataene raskere og mer robust.

Et skjellmateriale av rømt oppdrettslaks fra elver på Vestlandet er samlet inn og analysert av Rådgivende Biologer ved Kurt Urdal og Harald Sægrov. Det finske forskningsinstituttet LUKE har bidratt med skjell av rømt oppdrettslaks fra grenseelvene Tana og Neiden. Øvrige skjell er innsamlet av NINA.

Forskningssjef Kevin Glover ved Havforskningsinstituttet bidro med DNA-materiale samt råd vedrørende design av SNP-arrayet.

Referansegruppen for dette prosjektet har bestått av Forskningssjef Thomas Moen (AquaGen) og Genomics specialist Matthew Baranski (Mowi), og den har gitt meget nyttige bidrag til utforming av rapporten. Selv om prosjektgruppen står ansvarlig for metodikk, resultater tolkning og konklusjoner, har vi likevel synliggjort de viktigste gjenstående faglige uenighetene i fotnoter for å legge til rette for eventuell fremtidig diskusjon med næringen.

Kjell Maroni har vært den ansvarlige hos FHF for dette prosjektet.

2.0 Problemstilling og formål

2.1 Prosjektets effektmål

1. Prosjektet dokumenterer næringens interesse i å redusere problemet med rømt oppdrettslaks, noe som vil kunne bidra til å redusere spenningen mellom villaksinteressene, miljøforvaltningen og næringen.

2. Påvisning av at rømt oppdrettslaks fanget i elv er genetisk mer lik villaks enn annen oppdrettslaks vil gi avlsselskapene et første grunnlag for å vurdere om denne informasjonen lar seg utnytte i praksis, og hva som trengs av ytterligere informasjon før den eventuelt kan innlemmes i avlsarbeidet.

3. Identifisering av genetisk variasjon av mulig betydning for å vandre opp i ferskvann vil kunne muliggjøre gjennomføring av relativt kortvarige eksperimentelle verifiseringsforsøk, for eksempel kontrollerte smolt- og voksenfisk-utsettinger. Resultatene fra slike forsøk vil være et viktig bidrag til operasjonalisering av kunnskapen.

4. Publisering av detaljerte positive resultater vil sannsynligvis medføre at ingen kan komme senere og påberope seg IPR på oppdrettslaks som ikke lenger er i stand til å vende tilbake til ferskvann.

2.2 Prosjektets resultatmål

Prosjektets resultatmål var å avklare:

1. Om det finnes klare genetiske forskjeller mellom den generelle oppdrettspopulasjonen og oppdrettslaks fanget i norske lakseelver,
2. Om det er forskjeller, er oppdrettslaks fanget i elv mer lik villaks enn annen oppdrettslaks,
3. Kan de genetiske forskjellene kobles til biologiske mekanismer som underligger laksens evne til å vende tilbake til ferskvann.

3.0 Prosjektgjennomføring

3.1 Metodikk

Dersom oppdrettslaksen gjennom avlsarbeidet har blitt selektert for egenskaper som indirekte har forårsaket en sterk nedtoning av evnen til å søke tilbake til ferskvann, vil rømte individer fra en gitt avlspopulasjon (AquaGen, SalmoBreed, Mowi, Rauma) som fremdeles går opp i ferskvann representere en kontrastpopulasjon til denne avlspopulasjonen. Dette innebærer at disse fire kontrastpopulasjonene vil alle måtte dele en genetisk signatur med villaksen som ikke forefinnes i majoriteten av individer fra avlspopulasjonene. Innsamling av biologisk materiale, genotyping og påfølgende analyse ble rigget for å bekrefte eller avkrefte eksistensen til en slik genetisk signatur, og eventuelt dokumentere at den genetiske signaturen involverer gener assosiert med biologi som kan knyttes til *ferskvannsoppvandringsevne*. Det vil si **gener assosiert med biologiske mekanismer av sannsynlig betydning for evne til å unngå predasjon (tap av frykt for predatorer eller evne til flukt/unnvikelse), evne til å fange byttedyr i sjøfasen etter rømming, evne til å navigere i sjø over korte og lange distanser, og evne til å søke opp i ferskvann basert på luktstimuli**. Vår forventning var at vi ville finne genetiske signaturer som i stor grad er knyttet til nevrobiologi, lukteevne, hukommelse og læring.

Grunnen til å etablere fire kontrastpopulasjoner som stammer fra fire ulike avlspopulasjoner, var at dette gir et langt bedre grunnlag for å bekrefte eller avkrefte prosjektets hypotese samt avdekke mulig kausal genetisk variasjon, enn om en kun hadde betraktet rømte individer fanget i elv som en gruppe og sammenlignet med villaks og kun en avlsstamme.

3.2 Gjennomføring av prosjektet

3.2.1 Utvelgelse av 2000 rømte oppdrettslaks fanget i elv og ekstraksjon av DNA fra disse (Delaktiviteter 1 og 2)

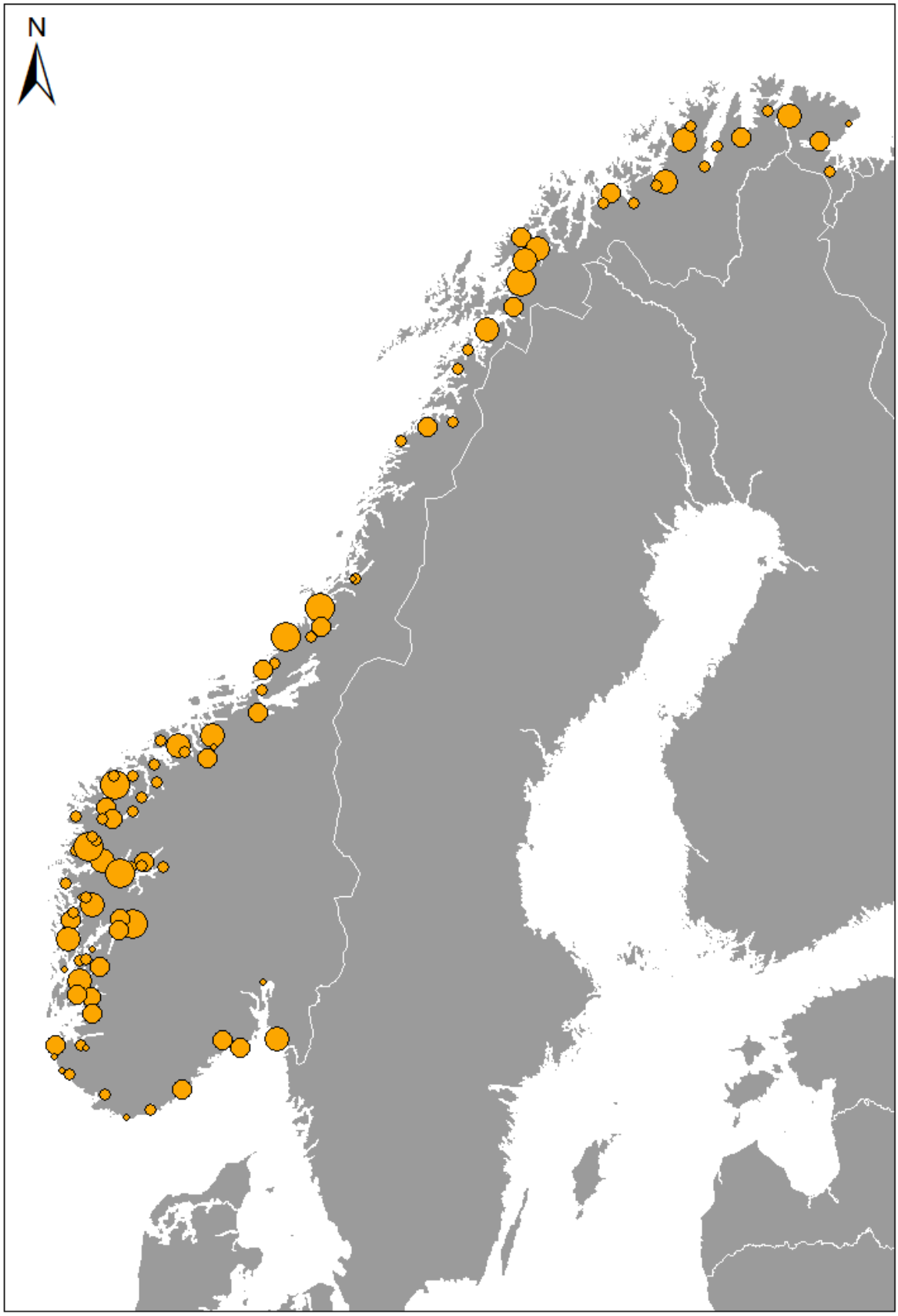
NINA har i samarbeid med Rådgivende Biologer, Veterinærinstituttet og det finske forskningsinstituttet LUKE gjennomgått skjellprøver fra mer enn 2100 rømte oppdrettslaks fra sportsfiske i elv (sommer) og fra organisert høstfiske (inkludert stamfiske etter villaks) de siste årene. I overkant av 100 av disse ga ikke god nok DNA-kvalitet for genotyping. En oversikt over de 1980 analyserte individene av rømt oppdrettslaks er gitt i Vedlegg 1 (som Excel-fil), og den geografiske fordelingen er vist i Figur 1. Om lag 360 av disse kommer fra elver der villaksen tilhører en mer østlig fylogenetisk gruppe (Barents-Kvitsjøen-gruppen) enn laksen i resten av Norge (som tilhører den Østatlantiske fylogenetiske gruppen). DNA ble ekstrahert fra skjellene ved bruk av Qiagen tissue ekstraksjonskit.

Vi la vekt på å bruke rømt oppdrettslaks fanget i elv gjennom et forholdsvis begrenset tidsrom. Da vi ønsket at oppdrettslaksen fanget i elv i størst mulig grad skulle kunne representere deres respektive avlspopulasjoner valgte vi stikkprøver noen år etter at prøver av oppdrettslaks fra de ulike avlsselskapene ble samlet inn, det vil si fangstårene 2015-2016 og dernest 2014. I noen elver – bl.a. Tanavassdraget der prøvene er innhentet fra det finske forskningsinstituttet LUKE – er materialet vårt fra et lengre tidsrom og stort sett eldre enn andre stikkprøver.

De 1980 prøvene er samlet inn fra 98 elver fra Glomma i sørøst (nær grensen til Sverige) til Neiden i nordøst, som vi deler med Finland. Vi har lagt vekt på å ha 20 eller flere prøver av rømt oppdrettslaks fra elver der vi er kjent med at det jevnlig søker opp rømt oppdrettslaks for å gyte, og har om lag 100 prøver fra noen av elvene. Vi har ekskludert fra materialet individer som ble fanget som umodne, dvs individer som ikke skal gyte det året de ble fanget. Vi har også lagt vekt på å klassifisere den rømte oppdrettslaksen med tanke på om de er nyrømt eller om de har hatt én eller flere vintre i sjøen etter rømming. Den siste gruppen er spesielt interessant siden disse individene har hatt et opphold ute i storhavet før de har søkt tilbake til elv for å gyte.

3.2.2 Utvikling av nytt 60K SNP-array **(Delaktivitet 3)**

CIGENE, NMBU, er ledende i Norge på utvikling av SNP-array for en rekke arter, og konstruksjon av nytt 60K SNP-array for å genotype de utvalgte rømte oppdrettslaksene og ytterligere individer fra SalmoBreed, Mowi og Rauma fulgte senterets etablerte protokoller. AquaGen ga i forkant av prosjektet tilslutning til bruk av sitt 220K SNP-array som utgangspunkt for assay-design. Premissene for denne bruken har blitt nedtegnet i en egen avtale. SNP-arrayet ble designet slik at det i tillegg til å oppfylle behovene for dette prosjektet også ville ha stor bruksverdi både for industrien og forskningsmiljøene (spesifikasjon gitt nedenfor). Av 74,600 utvalgte SNPer som ble sendt til Thermo Fisher Scientific (tidligere Affymetrix) for design av genotyping assays, endte vi opp med et endelig markørsett bestående av 60,316 SNPer.



Figur 1: Geografisk fordeling av 1980 rømte oppdrettslaks fanget i elv i Norge, hovedsakelig i årene 2014-2016. Størrelsen på symbolene angir om materialet fra elva består av fra 1 (minst symbol), 2-10, 11-30, 31-80, eller flere enn 80 (størst symbol) skjellprøver av rømt oppdrettslaks. Illustrasjon: Anders Foldvik, NINA.

3.2.3 Genotyping av rømt oppdrettslaks fanget i elv og individer fra oppdrettslakspopulasjonene (Delaktiviteter 4 og 5)

Det nye 60K-arrayet ble brukt til å genotype de utvalgte oppdrettslaksene fanget i elv og omlag 120 individer fra henholdsvis SalmoBreed, Mowi og Rauma. Genotypingen ble utført av CIGENE (www.cigene.no) i henhold til standard protokoller for Axiom teknologi utviklet av Thermo Fisher Scientific (www.thermofisher.com/no/en/home.html).

3.2.4 Tilordning av oppdrettslaks fanget i elv til avlsstamme (Delaktivitet 6)

De genotypede rømte oppdrettsindividene fanget i elv ble tilordnet opphavspopulasjon (AquaGen, Mowi, Rauma, SalmoBreed) ved hjelp av metoder tidligere utviklet ved NINA. Materialet ble også analysert med NINA-utviklede metoder som skiller mellom villaks og oppdrettslaks. For individer der klassifiseringen med gentestene var tvetydige ble vekstmønsteret til de aktuelle skjellprøvene gjennomgått på nytt. Skjellmønstrene til villaks og oppdrettslaks er så distinkt forskjellige for erfarne skjellesere at de i de aller fleste tilfeller kan brukes som fasit. Om skjellprøvene ikke var entydige ble disse individene ekskludert for sikkerhets skyld.

3.2.5 Sammenligning av genetiske signaturer (Delaktiviteter 7 og 8)

Analysearbeidet ble i all hovedsak gjort med egenutviklet programvare skrevet i Python og ved bruk av utviklingsverktøyet JupyterLab. Motivasjonen for dette var at datamaterialet hadde en slik beskaffenhet at informasjonsverdien i det ikke lot seg ekstrahere ved hjelp av eksisterende bioinformatisk programvare for påvisning av genetiske assosiasjoner (inklusive state-of-the-art Genome Wide Association Studies (GWAS) programvare).

Det ble utviklet fem ulike analyseverktøy i JupyterLab:

FiloPyth_Prune: Scriptet identifiserer alle SNPer hvor forskjellen mellom allelfrekvens til de fire kontrastpopulasjonene (dvs rømt oppdrettslaks fanget i elv tilordnet henholdsvis AquaGen, SalmoBreed, Mowi og Rauma) og villaks er lavere enn en gitt terskelverdi. Av disse SNPene velger scriptet ut de som er felles for alle fire kontrastpopulasjonene. Av dette snittet plukker det ut de SNPene hvor allelfrekvensforskjellen mellom hver avlspopulasjon og villaks er større enn en gitt terskelverdi.

FiloPyth_Combine: Scriptet importerer SNP-settet generert av det foregående scriptet (FiloPyth_Prune) og lager alle mulige distinkte 2- og 3-SNP-kombinasjoner av disse. Deretter tar scriptet hver enkelt kombinasjon og bestemmer frekvensen av alle mulige kombinasjoner av genotyper (henholdsvis 9 og 27) hos villaks, i en gitt avlspopulasjon og i rømt oppdrettslaks fanget i elv som stammer fra denne avlspopulasjonen. Det produserer så en score-tabell som viser hvilke genotypekombinasjoner for hver enkelt SNP-kombinasjon som er tilstede i høyest frekvens i de tre populasjonene. For å redusere regnetiden ble 2-SNP-kombinasjonsresultatene brukt som utgangspunkt for 3-SNP-kombinasjonskjøringene.

FiloPyth_Cover: Scriptet importerer data generert av det foregående scriptet (FiloPyth_Combine) og lager alle mulige 2- til 8-elementkombinasjoner av de mest lovende 3-SNP-kombinasjonene. I 2-elementtilfellet får en da følgende Boolske funksjon: {Genotype_SNP1 AND Genotype_SNP2 AND Genotype_SNP3} OR { Genotype_SNP4 AND Genotype_SNP5 AND Genotype_SNP6}. Denne Boolske funksjonen som kan anta verdien Sann eller Falsk og kan betraktes å være en førsteordens approksimering av den sanne logiske strukturen til genotypesettene, blir så testet på alle individer i alle 9 populasjonene for å bestemme hvor stor andel av hver enkelt populasjon som har en genotypesammensetning som gir verdien Sann. Tilsvarende for 4-8-elementkombinasjonen. For å redusere regnetiden ble også her 2-elementkombinasjonsresultatene brukt som utgangspunkt for 4-8-element kombinasjonskjøringene.

FiloPyth_Annotate: Scriptet importerer de mest lovende kausale kandidatsettene fra FiloPyth_Cover og annoterer hver SNP ved hjelp av en database utviklet av CIGENE som angir kromosom, kromosomposisjon,

hvilken type SNP det er i henhold til design av 60K-arrayet, hvilken genetisk funksjon SNPen har, og identiteten til SNPen i NCBI sin database som muliggjør bruk av komparativ funksjonell genomisk informasjon. For hver SNP blir det deretter søkt etter gener i området rundt SNPen ved hjelp av en database tidligere utviklet av CIGENE som angir alle gener ±50Kb av alle SNPer på 220K-arrayet til AquaGen. Den biologiske funksjonen til disse genene blir så identifisert ved hjelp av database- og litteratursøk.

FiloPyth_Boolean: Dette scriptet søker å avdekke ytterligere mønstre i datasettene for i enda sterkere grad å avdekke kausale strukturer og genotypekoblinger. Scriptet importerer 4-8-elementkombinasjonene fra FiloPyth_Cover som gir den høyeste dekningen for villaks og rømt oppdrettslaks og lav dekning av avlspopulasjonene, samt informasjon fra FiloPyth_Annotate. Det anvender Boolske analyseteknikker som brukes i forbindelse med konstruksjon av mikroprosessorer, og det anvender teknikker for å sammenligne tekststrenger (såkalt Levenshtein-metrikk).

For å kvalitetssikre den egenutviklede programvaren anvendte vi også en eksisterende programvarepakke for logisk regresjon*:* logicFS: Identification of SNP Interactions (link: https://rdrr.io/bioc/logicFS/) utviklet i programspråket R. Denne pakken finner også Boolske relasjoner i SNP-sett, men ikke på samme detaljeringsnivå som det vi mener er nødvendig. Men de strukturene den finner vil måtte overensstemme med en undermengde av de strukturene vår programvare avdekker.

De fleste kjøringene ble gjort på regneressurser velvillig stilt til disposisjon av UNINETT Sigma2 AS, den nasjonale infrastrukturen for beregningsvitenskap i Norge.

4.0 Oppnådde resultater, diskusjon og konklusjon

4.1 Detaljert oversikt over oppnådde resultater

4.1.1 Karakterisering av nytt 60K SNP-array

60K SNP-arrayet ble designet i fem steg:

I **Steg 1** valgte vi ut ca. 15000 SNPer fra analyse av 220K SNP-array data på villaks og oppdrettslaks:

- SNPer i selection sweeps i oppdrett basert på tidligere analyser i CIGENE
- SNPer koblet til ulike miljøforhold i villaks basert på tidligere analyser av CIGENE i NFR-prosjektet QuantEscape.
- SNPer med store forskjeller i allelfrekvens mellom villaks og oppdrettslaks (AquaGen, Mowi, Rauma og SalmoBreed) basert på analyser i QuantEscape.
- SNPer med store forskjeller i allelfrekvens i ulike oppdrettspopulasjoner (AquaGen, Mowi, Rauma og SalmoBreed) basert på analyser i QuantEscape.

I **Steg 2** la vi til et lite antall SNPer tidligere brukt til å kvantifisere bidrag av oppdrettsfisk i villaks (fra NINA, publisert av Karlsson et al., 2011, og Havforskningsinstituttet, upublisert) samt noen få SNPer fra en studie publisert av en forskergruppe i Skottland (totalt 145 stk.).

I **Steg 3** la vi til ca. 3500 SNPer med sannsynlig funksjonell effekt (snpEff), med en minimum-frekvens i villaks og en viss forskjell i allelfrekvens mellom villaks og oppdrett. Data for å identifisere disse SNPene kom fra resekvensering av mer enn 400 villaks (innsamlet av NINA og Rådgivende Biologer fra Østfold til Finnmark) og 48 AquaGen fisk i AquaGenome Project, finansiert av NFR og AquaGen.

I **Steg 4** fylte vi inn med mest mulig informative SNPer fra 220K for å sikre best mulig fordeling av SNPer i genomet (opp til 50000 SNPer).

I **Steg 5** fylte vi inn 'hull' i genomet ved bruk av mest mulig informative SNPer fra resekvensering i AquaGenome Project (opp til 74600 SNPer).

I tillegg til ordinære SNPer ble det det lagt til prober på arrayet som muliggjør bestemmelse av genetisk kjønn til fisken (fravær eller tilstedeværelse av kjønnsbestemmende lokus sdY).

Av de 74600 SNPene som ble sendt til Thermo Fisher Scientific for design av genotyping-assays, endte vi opp med et endelig markørsett bestående av 60316 SNPer. Av disse er det omlag 12000 SNPer som dette prosjektet ikke kunne nyttiggjøre seg grunnet behovet for sammenligning med villaks og oppdrettslaks som tidligere hadde blitt genotypet med 220K SNP-arrayet, men som øker den fremtidige bruksverdien av arrayet til andre formål. 48075 SNPer av de 60316 på arrayet ble brukt for den videre analysen.

4.1.2 Validering av nytt 60K SNP-array

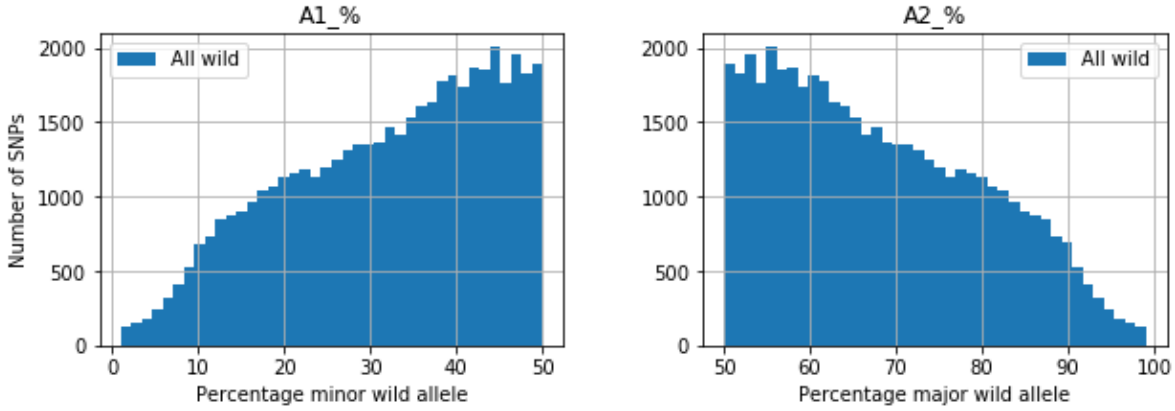
Vi testet 32 individer som tidligere hadde blitt genotypet på 220K-arrayet på det nye 60K-arrayet. Kun 76 SNPer hadde en diskordansverdi større enn hva som vanligvis settes som terskel (0.1). Da dette validerinsgresultatet av ulike årsaker kom noe sent, ble disse SNPene spesifikt sjekket i etterkant av analysene. Liste over disse SNPene er gitt i Vedlegg 2. Distribusjonen av diskordansverdier viser at data fra det nye arrayet i svært stor grad er overensstemmende med data fra 220K-arrayet.

4.1.3 Tilordning av rømt oppdrettslaks fanget i elv til avlspopulasjon

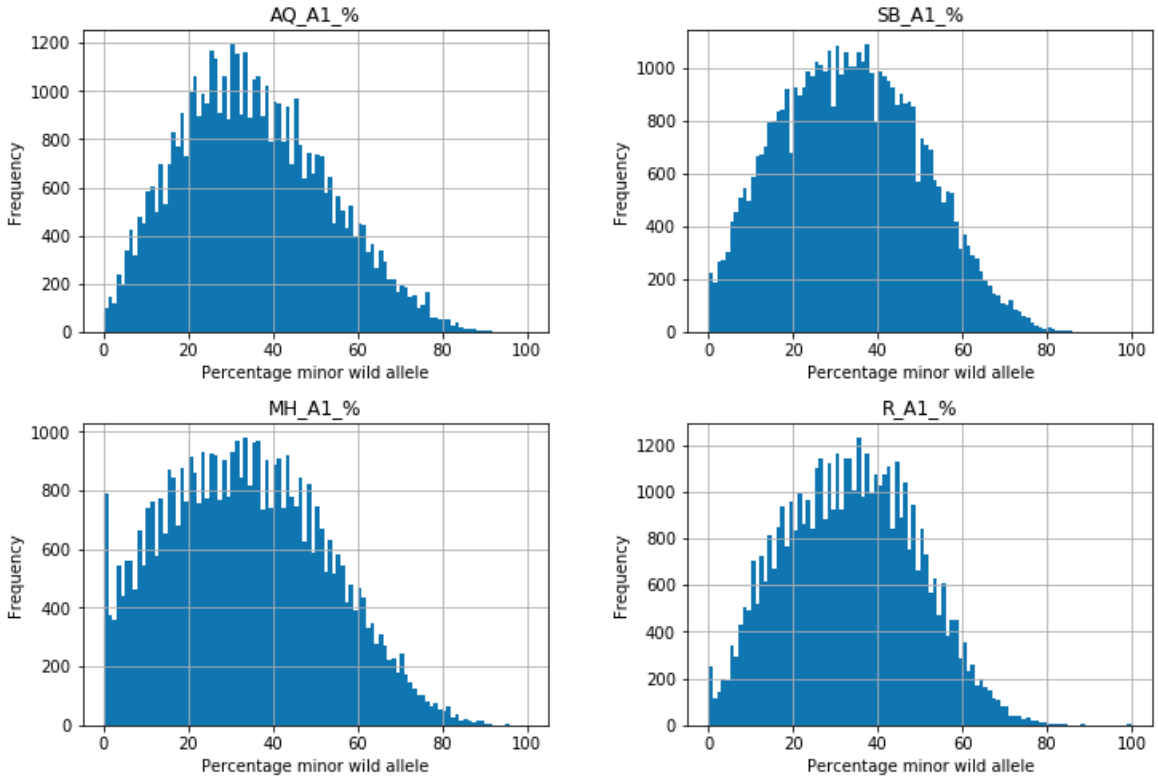
Vellykket genotyping ble gjennomført med 60K SNP-arrayet for 1888 rømte oppdrettsindivider fanget i elv. Totalt ble 1920 individer genotypet, men 32 kunne ikke brukes på grunn av for dårlig DNA-kvalitet. Totalt 1775 av de opprinnelige 1980 genotypede oppdrettsindividene fanget i elv har blitt verifisert som reelle basert på gentester og skjellprøvemønster. Disse fordeler seg som følger mellom de ulike avlspopulasjonene: AquaGen: 536; SalmoBreed: 542; Mowi: 283; Rauma: 414

4.1.4 Analyseresultater

4.1.4.1 Allelfrekvensfordelinger i avlspopulasjonene

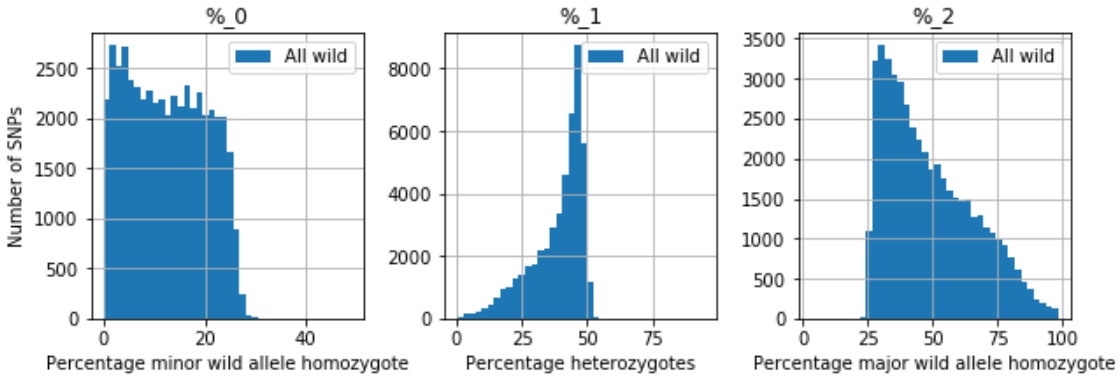


Figur 2: Prosentvis fordeling av SNP-allelet med henholdsvis lavest (A1) og høyest (A2) frekvens i villaks for alle 48075 SNPer som ble anvendt på 60K arrayet. Denne allel-benevningen ble konsekvent brukt på alle de fire avlspopulasjonene og på rømt-populasjonen for å kunne sammenligne populasjonene både med hensyn til allelfrekvensfordelinger og genotypefrekvensfordelinger.

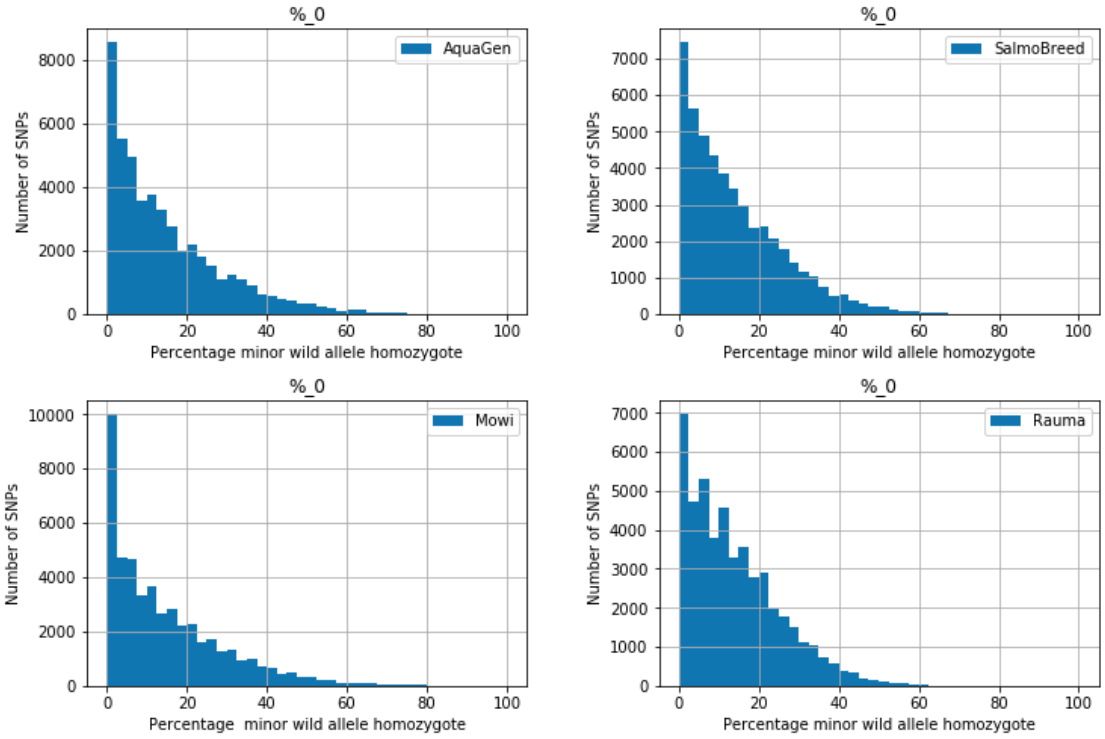


Figur 3: Fordelingen av A1 i alle fire avlspopulasjonene (AQ = AquaGen, SB = SalmoBreed, MH = Mowi, R = Rauma).

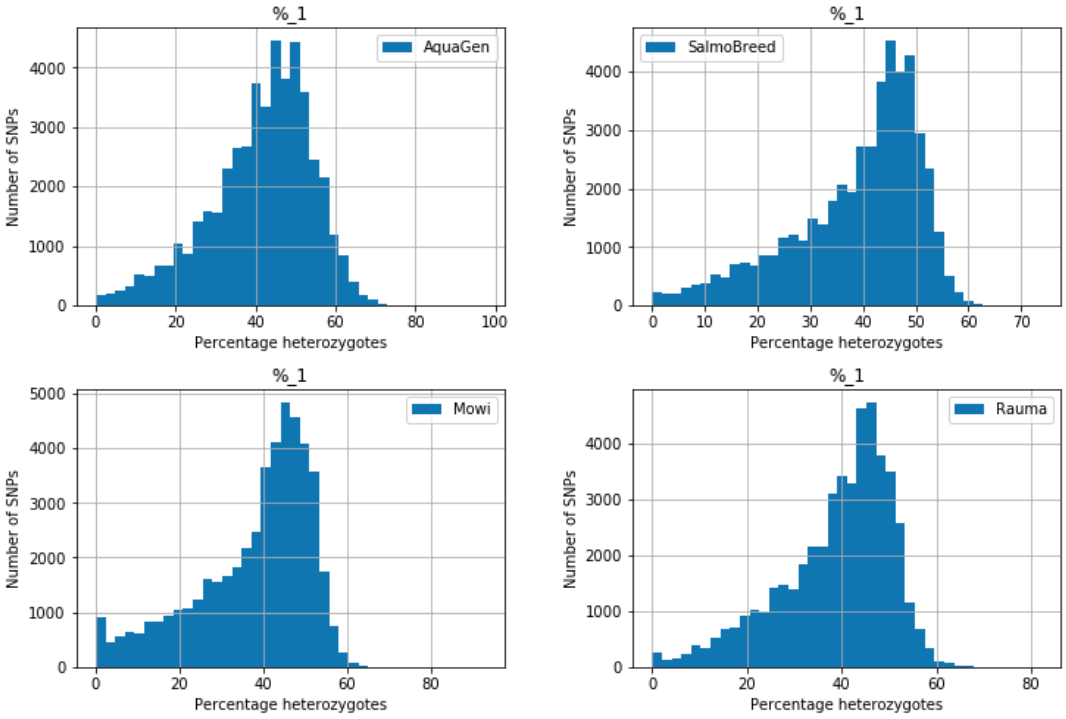
4.1.4.2 Genotypefrekvensfordelinger i avlspopulasjonene



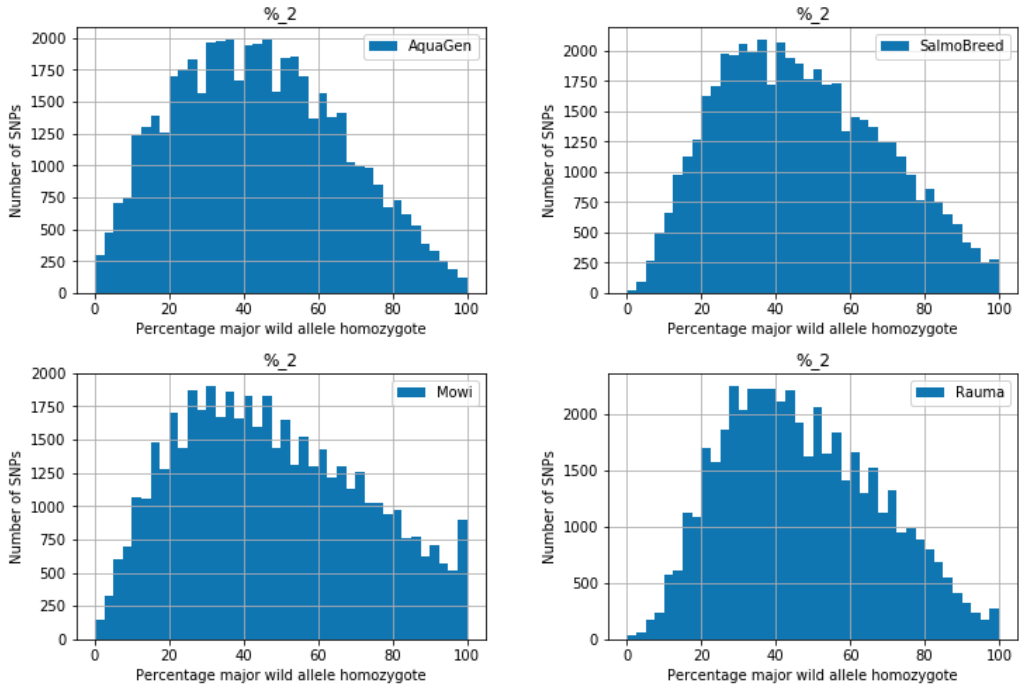
Figur 4: Fordelingen av genotypene A1A1 (0), A1A2 (1) og A2A2 (2) over alle 48075 SNPer i villakspopulasjonen.



Figur 5: Fordelingen av den homozygote genotypen A1A1 over alle 48075 SNPer i de fire avlspopulasjonene.



Figur 6: Fordelingen av den heterozygote genotypen A1A2 over alle 48075 SNPer i de fire avlspopulasjonene.



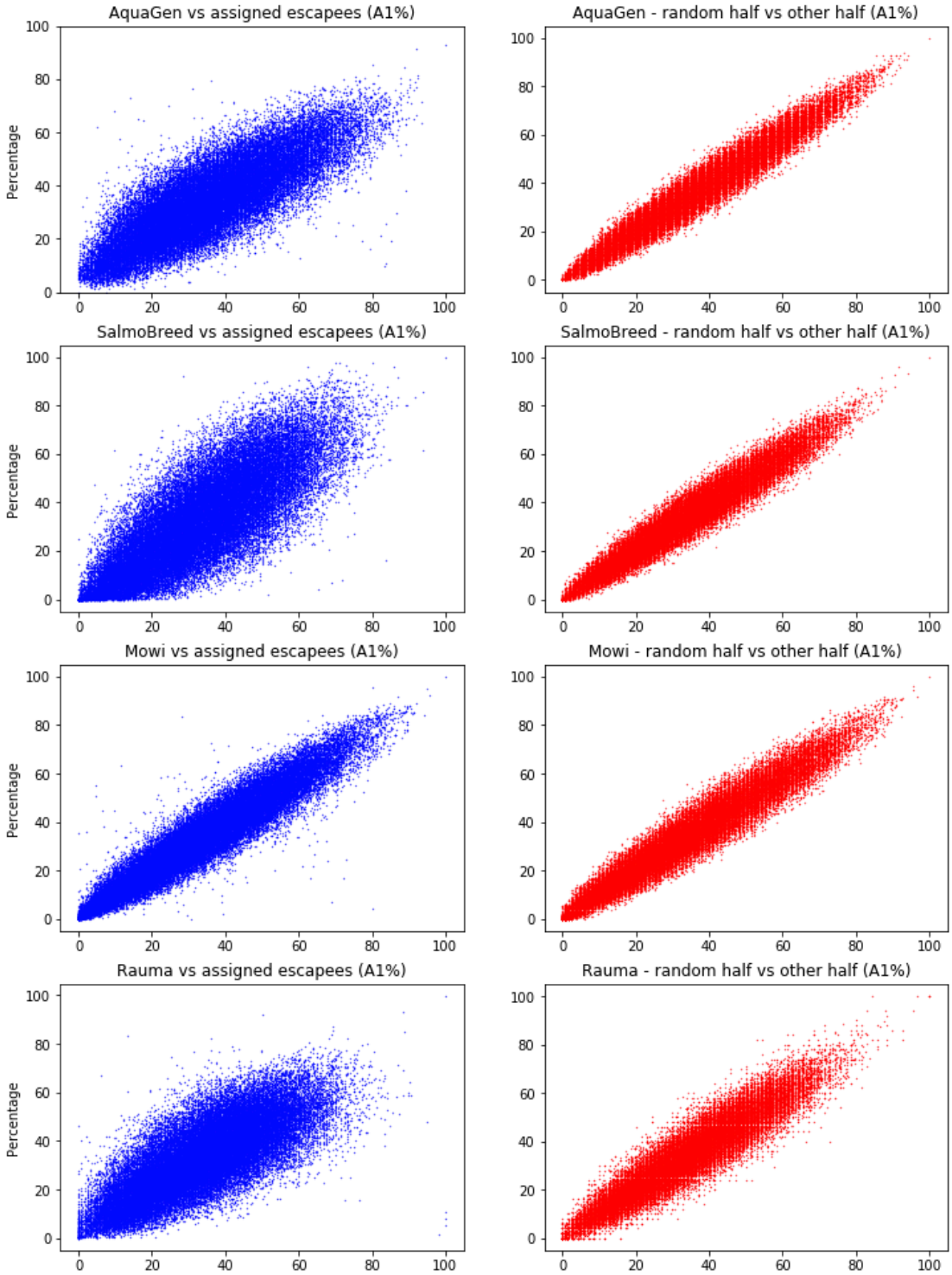
Figur 7: Fordelingen av den homozygote genotypen A2A2 over alle 48075 SNPer i de fire avlspopulasjonene.

4.1.4.3 Genetisk representativitet til rømt oppdrettslaks fanget i elv i forhold til opphavspopulasjon
Den mulige praktiske nytten av dette prosjektet står og faller på at rømt oppdrettslaks fanget i elv ikke er genetisk representativ for den avlspopulasjonen de kommer fra. En analyse av denne representativiteten er avhengig av to premisser. Det første er at de individene vi har fått fra avlsselskapene er representative for de enkelte avlskjernene for den årgangen de ble plukket fra. Selv om dette utvalget har skjedd på forbilledlig vis kan det likevel være utfordrende å sikre representativitet. Det andre er at avlskjerneindividene er rimelig representative for de populasjonene som de facto befinner seg i merdene, det vil si at den genetiske forskjellen mellom avlskjerneindivider og produksjonsindivider ikke er svært stor. Referansegruppen påpeker at de fleste eller alle avlsselskaper har minst én runde med seleksjon mellom avlspopulasjon og salgsrogn, fisken i en enkeltmerd representerer bare en liten del av den totale genetiske variasjonen i avlspopulasjonen, og at mor og far til produksjonsfisken tas ofte fra separate grupper med fisk, med ulike slektskap til avlskjernene. I tillegg kommer at produksjonsfisk kan tilhøre årsklasser som er yngre eller eldre enn årsklassene som dominerer i avlskjerne-referansene brukt i prosjektet. Det er derfor betydelig usikkerhet knyttet til hvor stor den genetiske distansen mellom avlspopulasjonene og de respektive produksjonspopulasjonene var for de angjeldende årene, og slike data er ikke tilgjengelige.

Da det er lite sannsynlig at en bedre representativitet ville kunne snu en negativ konklusjon til en positiv konklusjon, har utgangspunktet vårt for analysearbeidet under vært at om vi ikke kunne påvise forskjeller mellom rømt oppdrettslaks og deres respektive avlspopulasjoner med de dataene vi hadde til rådighet, så ville vi konkludere med at det ikke er grunnlag for å gå videre med en oppfølgingsstudie som inkluderer spesifikk adressering av disse to representativitetsspørsmålene.

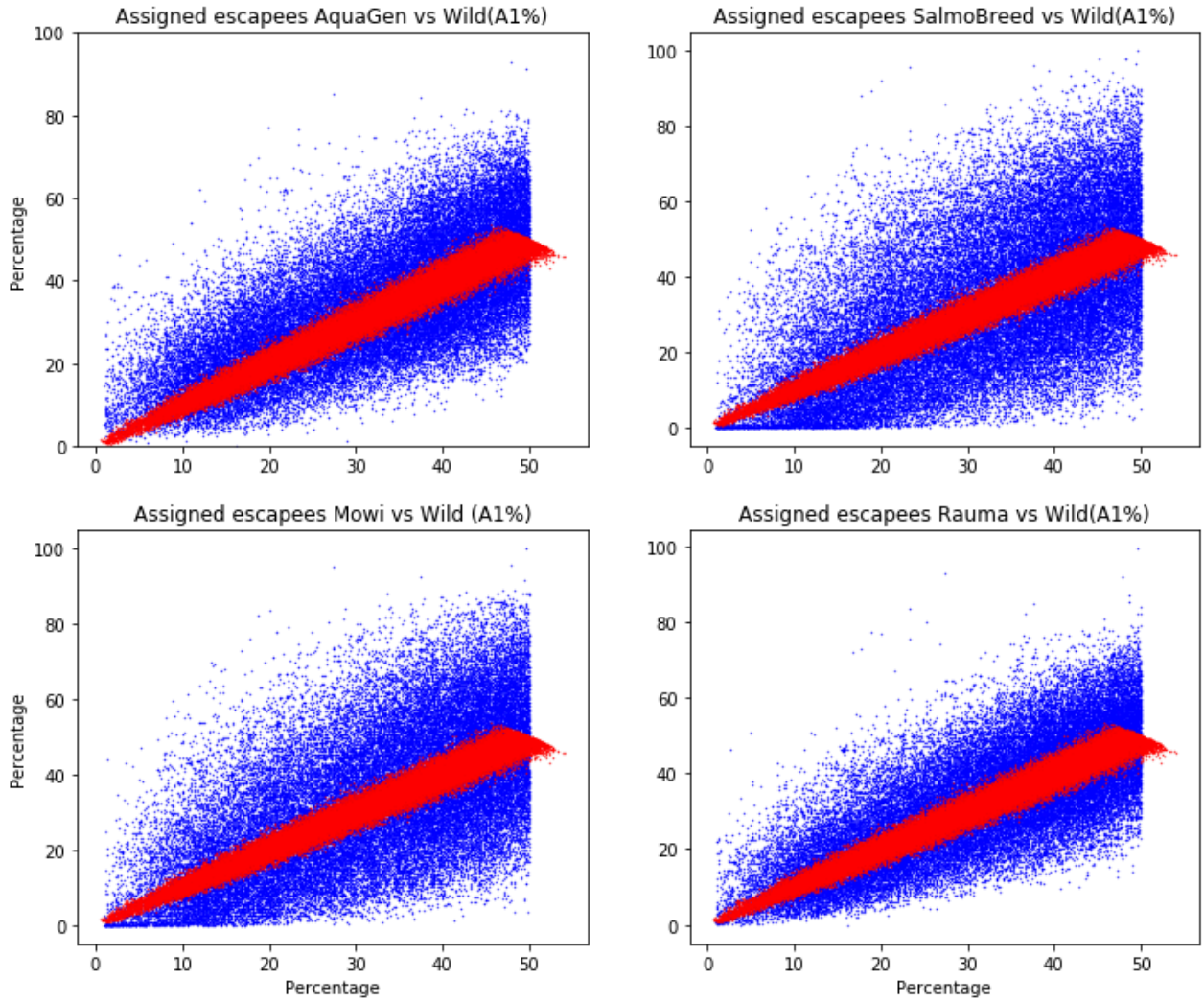
Venstre kolonne i Figur 8 under viser scatterplots for hvordan frekvensen av A1 for en gitt avlspopulasjon og rømte individer fra denne avlspopulasjonen samsvarer med hverandre over alle 48075 SNPer. Den høyre kolonnen viser samme type plot, men her er hver avlspopulasjon splittet i to tilfeldige halvdeler som er sammenlignet med hverandre. En ser at det er betydelig mindre samvariasjon mellom rømt oppdrettslaks fanget i elv og respektiv avlspopulasjon enn mellom to tilfeldige halvdeler av hver avlspopulasjon. Selv om det er klar forskjell mellom plottene for Mowi, så skiller denne populasjonen seg noe ut ved at forskjellen

ikke er så distinkt som for de tre andre tilfellene. Under forutsetning av at avlspopulasjonene er rimelig representative i begge betydninger (jfr. reservasjonene over) indikerer disse plottene at rømt oppdrettslaks fanget i elv skiller seg genetisk fra opphavsstammene. Men det er verdt å bemerke at antall individer i de enkelte avlspopulasjonene er forholdsvis lavt, og en utvidelse av antallet vil kunne øke spredningen slik at forskjellene ikke blir så distinkte. Kun ytterligere data vil kunne bekrefte eller avkrefte dette.



Figur 8: Scatterplots av frekvensfordelingen til A1 over alle 48075 SNPer for rømtpopulasjon mot opphavspopulasjon (venstre kolonne) og en tilfeldig utvalgt halvdel av hver avlspopulasjon mot den andre halvdelen (høyre kolonne). Se tekst for ytterligere forklaring.

For ytterligere å sjekke kvaliteten på skjellesingsarbeidet og dermed forsikre oss om at den overveldende majoriteten av laks fanget i elv virkelig var rømt oppdrettslaks og ikke feilklassifiserte villaks, gjorde vi et plot analogt til Figur 8, men nå med rømt oppdrettslaks plottet mot villaks (Figur 9). I dette tilfellet er sammenligningen basert på nullhypotesen om at rømt oppdrettslaks fanget i elv er i virkeligheten villaks.



Figur 9: Scatterplots av frekvensfordelingen til A1 over alle 48075 SNPer for rømtpopulasjon mot villaks (i blått) og en tilfeldig utvalgt halvdel av villakspopulasjonen mot den andre halvdelen (innfelt rødt). Se tekst for ytterligere forklaring. At det røde plottet stikker litt ut av det blå plottet er kun et artefakt og reflekterer ikke at noen SNPer har en A1-allelfrekvens høyere enn 50 % i villakspopulasjonen.

Vi ser at variasjonen mellom rømtlaks og villaks er mye større enn variasjonen mellom to tilfeldig utvalgte halvdeler av villakspopulasjonen for alle rømtpopulasjonene. Det utelukker ikke at vi fremdeles kan ha gjort noen feilklassifiseringer, men plottene tyder på at de i tilfelle er så få at de ikke vil påvirke resultatene nevneverdig.

Til tross for usikkerhetene knyttet til Figur 8 fant vi mønstrene såpass overbevisende at det motiverte oss til, med utgangspunkt i det materialet som var tilgjengelig, å gjøre noen mer inngående allelfrekvensanalyser. Analysene som følger er basert på en forestilling om at avlsarbeidet, enten ved direkte seleksjon eller indirekte på grunn av sterk koblingsulikevekt med genvarianter under direkte seleksjon, har økt frekvensen av genvarianter som i utgangspunktet er i lav frekvens i villakspopulasjonen

og som er dysfunksjonelle for ferskvannsoppvandringsevnen. Men resultatene er ikke avhengige av gyldigheten til denne forestillingen.

Dersom rømt oppdrettslaks fanget i elv ikke er mer lik villaks enn opphavspopulasjonen sin er følgende relasjon gyldig:

$$\left\{\bigcap_{i=1}^{4}\left\{\left|f_{A1}^{k}(E_i)-f_{A1}^{k}(W)\right|<\alpha_1 \,\&\, f_{A1}^{k}(E_i)<\alpha_2 \,\&\, f_{A1}^{k}(W)<\alpha_2, \forall k\in S\right\}\right\}$$
$$\approx\left\{\bigcap_{i=1}^{4}\left\{\left|f_{A1}^{k}(A_i)-f_{A1}^{k}(W)\right|<\alpha_1 \,\&\, f_{A1}^{k}(A_i)<\alpha_2 \,\&\, f_{A1}^{k}(W)<\alpha_2, \forall k\in S\right\}\right\}$$

Her står E_i for rømt oppdrettslaks fanget i elv fra stamme *i*, A_i står for den respektive avlspopulasjonen, W står for villakspopulasjonen, S står for SNP-settet brukt (48075), f_{A1}^{k} står for frekvensen av A1 for SNP nummer *k*, og α_1 og α_2 står henholdsvis for hvor lik E_i eller A_i må være villaks og hva den høyeste allelfrekvensverdien for A1 som er med i utvalget. Med α_1 = 0.1 og α_2 = 0.2 får vi 2008 SNPer som er felles for rømtlakspopulasjonene (mengdesnittet av 5168 (AquaGen), 5201 (SalmoBreed), 5227 (Mowi) og 6142 (Rauma)) og 1628 SNPer som er felles for avlspopulasjonene (mengdesnittet av 3949 (AquaGen), 5641 (SalmoBreed), 5241 (Mowi) og 5911 (Rauma)). 1055 SNPer er felles for begge populasjonsgruppene. Det vil si nesten 50 % av alle SNPene felles for E-gruppen er ikke tilstede i A-gruppesnittet.

Da dette resultatet er ganske oppsiktsvekkende fant vi det legitimt å gjøre videre analyser som kunne muligens avklare om det kan forklares med manglende representativitet til avlspopulasjonsdataene våre eller stor genetisk distanse mellom avlspopulasjonene og produksjonspopulasjonene.

Dersom rømtlaksene, som resultatet over tilsier, er mer lik villaks enn bakgrunnen forventer en at denne andelen øker dersom en reduserer α_1. Setter vi α_1 = 0.07 og beholder alt annet likt har E-gruppen 985 felles-SNPer mens A-gruppen har 809 felles-SNPer. 385 SNPer er felles for begge gruppene, det vil si E-gruppen har 600 SNPer (>60 %) som ikke er tilstede i A-gruppens felles-SNPer. Med α_1 = 0.05 har E-gruppen 403 felles-SNPer mens A-gruppen har 328 felles-SNPer. 109 SNPer er felles for begge gruppene, det vil si 73 % av felles-SNPene i E-gruppen er ikke tilstede i A-gruppens felles-SNPer.

Om en fraviker kravet om likhet med villaks, men fremdeles fokuserer på kun de SNPene som har en A1-frekvens i villaks mindre enn 0.2, så har A-gruppen 1130 SNPer av disse felles hvor A1-frekvensen er større enn 0.2 (mengdesnittet av 4950 (AquaGen), 3228 (SalmoBreed), 3127 (Mowi) og 3132 (Rauma)), og E-gruppen har 1006 SNPer som er felles (mengdesnittet av 4075 (AquaGen), 3051 (SalmoBreed), 3200 (Mowi) og 3047 (Rauma)). De to gruppene har 718 felles-SNPer. Det vil si >70 % av SNPene i E-gruppen er felles med A-gruppen, mens i de tre analysene over var representasjonen henholdsvis 52 %, 39 % og 27 %. I dette tilfellet er derfor en betydelig større andel SNPer i E-gruppen felles med A-gruppen. Dette betyr at når vi fjerner føringen som krever at gruppene skal være lik villaks, og vi ser på de SNPene hvor A1 har økt i frekvens i forhold til villaks, så øker andelen felles-SNPer mellom E- og A-gruppen. Dette indikerer at E-gruppen har en undermengde av genotyper som primært skiller dem fra opphavspopulasjonene sine, og at denne undermengden er assosiert med likhet til villaks.

Antar man at den genetiske distansen mellom rømtpopulasjonene og produksjonspopulasjonene hvor rømtlaksene kommer fra er liten, impliserer resultatene at for SNPer med A1-frekvens < 0.2 i villaks så er produksjonspopulasjonene til de fire selskapene langt mere lik villaks enn sine respektive avlspopulasjoner. Ut fra følgende resonnement har vi har vanskelig for å se at et slikt synkront genetisk etterslep kan være mulig å bevare over tid basert på hvordan produksjonspopulasjonene etableres fra krysninger mellom stamfisk tatt fra avlspopulasjonene: For hver generasjon *k* >1 er det rimelig å anta at den genetiske distansen (ΔG) mellom produksjonspopulasjon (P) og avlspopulasjon (A) er mindre enn den genetiske

distansen mellom avlspopulasjonen og produksjonspopulasjonen generert fra forrige generasjons avlspopulasjon inntil oppstart, hvor distansen mellom A og P i første generasjon er mindre enn distansen mellom villaks og 1. generasjons produksjonspopulasjon:

$$\Delta G(A_{i,k},P_{i,k}) < \Delta G(A_{i,k},P_{i,k-1}) < .... < \Delta G(A_{i,k},P_{i,1}) < \Delta G(Wild,P_{i,1}), i = 1,..,4$$

Dette innebærer at den genetiske distansen mellom produksjonspopulasjon og villaks øker for hver generasjon for alle avlsstammene, og at den genetiske distansen mellom produksjonspopulasjonene fra de enkelte stammene øker over tid. Uansett hvilken metrikk man velger for å beskrive genetisk distanse er det derfor statistisk sett usannsynlig at produksjonspopulasjonene etter 12-15 generasjoner med avl skal samlet oppvise en mye større likhet med villaks enn avlspopulasjonene samlet. Selv om avlspopulasjons-dataene vi har brukt ikke er fullt ut representative, og selv om uttrykket over er en forenkling av hvordan avlsarbeidet har foregått, er de påviste forskjellene så store at avlspopulasjonsdataene må ha meget alvorlige mangler om dette resonnementet ikke skal være gyldig.

Om avlspopulasjonene ikke er representative for rømtfiskgruppene forventer vi at det er større slektskap innenfor avlspopulasjonene enn mellom avlspopulasjonene og rømtfiskgruppene. Det estimerte slektskapet (Wang, J. 2002, "An estimator for pairwise relatedness using molecular markers." *Genetics* 160, no. 3: 1203-1215) er:

Innenfor Mowi-kontroll:	0.0023
Innenfor Mowi-rømt:	-0.0155
Mellom gruppene:	-0.0213
Innenfor AquaGen-kontroll:	0.0627
Innenfor AquaGen-rømt:	-0.0009
Mellom gruppene:	-0.0153
Innenfor Rauma-kontroll:	-0.0224
Innenfor Rauma-rømt:	-0.0073
Mellom gruppene:	-0.0629
Innenfor SalmoBreed-kontroll:	-0.0346
Innenfor SalmoBreed -rømt:	0.0340
Mellom gruppene:	-0.0707

Det er generelt små forskjeller i slektskap (slektskap mellom helsøsken er 0.5), men vi observerer at slektskapet innenfor kontrollgruppene er noe høyere enn mellom kontrollgruppene og rømtfiskgruppene, spesielt for AquaGen.

Vi gjennomførte også en mer tradisjonell statistisk tilnærming for å teste om rømt oppdrettslaks fanget i elv var mer lik villaks enn oppdrettskontrollen ved å filtrere ut SNPer som hadde mindre enn 10 % A1-frekvens i villfisk. Dette ga et sett på 2174 SNPer. 339 av disse SNPene var på forhånd plukket ut for å være gode til skille oppdrettslaks fra villaks og ble derfor utelatt fra analysen, det resterende SNP-settet bestod derfor av 1835 SNPer. I tillegg ble SNPer der mer enn 10 individ manglet genotype ekskludert fra analysen. Endelig antall SNPer for hver analyse er gitt i Tabell 1. Forskjellen i gjennomsnittlig allelfrekvens mellom de fire rømtfiskgruppene og de fire kontrollgruppene (avlspopulasjonene) på dette SNP-settet ble testet statistisk for å undersøke om kontrollgruppene hadde en høyere allelfrekvens enn rømtfiskgruppene for disse SNP-ene (Tabell 1).

Tabell 1. Forskjell i gjennomsnittlig allelfrekvens mellom rømtfisk og avlspopulasjonene. Et negativt tall betyr at rømtfisken har lavere allelfrekvens og derfor er likere villfisken enn avlspopulasjonen.

Stamme	Estimat	Standardfeil	P-verdi	Antall SNP-er
Mowi	-0.00165	0.00067	0.0075	1718
AquaGen	-0.00732	0.00180	2.4×10^{-5}	1697
Rauma	-0.00959	0.00190	2.2×10^{-7}	1483
SalmoBreed	-0.00181	0.00084	0.0161	1676

I disse testene tok vi hensyn til slektskap gjennom å først estimere den gjennomsnittlige allelfrekvensen på tvers av disse SNPene for hvert individ. Vi estimerte så forskjellen i gjennomsnittlig allelfrekvens mellom gruppene i en «generalized least square» (GLS) modell der slektskapsmatrisen gir korrelasjonen mellom residualene i modellen. Med utgangspunkt i at vi brukte SNPer som hadde mindre enn 10 % A1-frekvens i villfisk, støtter resultatene forestillingen om at det finnes recessive alleler skadelige for ferskvanns-oppvandringsevnen i lav allelfrekvens i de ville populasjonene som har en høyere allelfrekvens i avlspopulasjonene på grunn av genetisk drift eller direkte eller indirekte seleksjon (se Seksjon 4.1.4.4 og Vedlegg 3 for underbygging av dette). Det er viktig å fremheve at vi ikke forventer store forskjeller i allelfrekvenser siden vi har tatt gjennomsnittet over mange SNPer og de fleste SNPene trolig ikke er under seleksjon i rømtfisken. For å validere resultatet gjennomførte vi samme analyse for 2174 tilfeldig valgte SNPer (repetert to ganger) (Tabell 2).

Tabell 2. Samme analyse som i Tabell 1, men på 1835 tilfeldig valgte SNP-er (repetert to ganger). SNP-er der mer enn 10 individ manglet genotype ble ekskludert fra analysen (antall brukt er gitt i siste kolonne).

Stamme	Estimat	Standardfeil	P-verdi	Antall SNP-er
Mowi (rep. 1)	-0.00033	0.00082	0.34	1646
(rep. 2)	-0.00080	0.00080	0.16	1645
AquaGen (rep. 1)	-0.00280	0.00144	0.026	1656
(rep. 2)	-0.00130	0.00138	0.17	1644
Rauma (rep. 1)	-0.00586	0.00170	0.00028	1573
(rep. 2)	-0.00353	0.00163	0.015	1579
SalmoBreed (rep. 1)	0.00046	0.00092	0.69	1593
(rep. 2)	-0.00130	0.00100	0.098	1602

I denne analysen observerer vi ikke de sterkt statistisk signifikante forskjellene som i Tabell 1. De estimerte forskjellene viser dog at rømtfisken i de fleste tilfeller er mer lik villfisken enn kontrollgruppen (altså negative estimater) og i tre tilfeller er forskjellen statistisk signifikant.

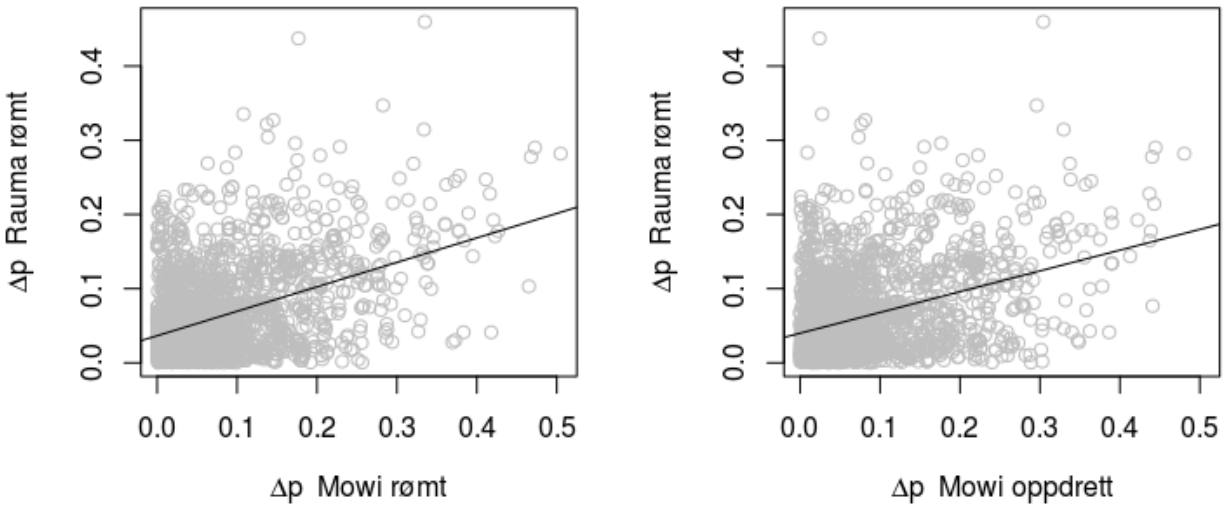
Gitt at rømtfisken er mer lik villfisk enn oppdrettsfisk vil vi forvente å finne en sterkere sammenheng i allelfrekvensforskjell til vill (Δp) mellom rømt-gruppene enn mellom rømt og oppdrett-kontroll. I 9 av 12 tilfeller var det en sterkere sammenheng (målt med R^2) for Δp mellom rømt og rømt enn mellom rømt og samsvarende oppdrettskontroll (Tabell 3).

Tabell 3. Sammenhengen mellom allelfrekvensforskjell til vill (Δp) for hver rømtfiskgruppe mot andre rømtfiskgrupper eller samsvarende oppdrettskontroll. Analysen ble gjort med en lineær modell med Δp for en rømtfiskgruppe som responsvariabel og en annen rømtfiskgruppe eller samsvarende oppdrettskontroll som forklaringsvariabel. Stigningstall med standardfeil (±SE) og R^2 er oppgitt for modellene. Modellene som best forklarer responsvariabelen er markert i grått. Det ene tilfellet der stigningstallet var brattere for modellen med lavest forklaringsgrad er markert med gult.

Respons (Bare rømt)	Forklaringsvariabel (Rømt eller kontroll)	Modell med rømt som forklaringsvariabel			Modell med kontroll som forklaringsvariabel		
		Stigningstall	±SE	R^2	Stigningstall	±SE	R^2

AquaGen	Mowi	0.044	±0.020	0.002	-0.018	±0.019	0.000
AquaGen	Rauma	0.302	±0.025	0.064	0.218	±0.024	0.038
AquaGen	SalmoBreed	0.053	±0.018	0.004	0.202	±0.022	0.036
Mowi	AquaGen	0.052	±0.023	0.002	0.007	±0.016	0.000
Mowi	Rauma	0.561	±0.025	0.185	0.597	±0.023	0.236
Mowi	SalmoBreed	0.511	±0.017	0.306	0.608	±0.021	0.271
Rauma	AquaGen	0.213	±0.017	0.064	0.077	±0.012	0.018
Rauma	Mowi	0.330	±0.015	0.185	0.280	±0.015	0.141
Rauma	SalmoBreed	0.281	±0.014	0.157	0.768	±0.010	0.735
SalmoBreed	AquaGen	0.074	±0.025	0.004	0.011	±0.017	0.000
SalmoBreed	Mowi	0.599	±0.019	0.306	0.524	±0.020	0.247
SalmoBreed	Rauma	0.559	±0.028	0.157	0.504	±0.026	0.143

Fra Tabell 3 ser vi også at stigningstallet som regel er brattere (8 av 12 tilfeller) mellom rømtgruppene enn mellom rømt og oppdrettskontroll. Dette betyr at rømtgruppene som regel er relativt mer like hverandre når allelfrekvensen er lik villfisken sammenlignet med den tilsvarende oppdrettskontrollen. Analysen for Rauma rømt mot Mowi rømt eller oppdrettskontroll er vist i Figur 10 som en illustrasjon.



Figur 10: Sammenheng mellom forskjell til rømt (Δp) mellom Rauma rømt og Mowi rømt eller Mowi oppdrettskontroll. (Tabell 2 oppgir stigningstall og R2 for sammenhengene.)

Det er verdt å fremheve at de forskjellene som ble avdekket med allelfrekvensanalysene over blir enda mer markante i genotypeanalysene som følger, så når en vurderer legitimiteten i påstanden om at rømtlaksene fanget i elv er mer lik villaks enn bakgrunnen må disse analyseresultatene også tas i betraktning. Men selv allelfrekvensanalysene alene gir grunn til å hevde at det er klare genetiske forskjeller mellom oppdrettslaks fanget i norske lakseelver og deres respektive avlspopulasjoner, og at oppdrettslaks fanget i elv har felles et betydelig antall SNP-loci hvor de er mer lik villaks enn hva som kan tilskrives tilfeldigheter. Størrelsen på forskjellene impliserer at den genetiske profilen til produksjonspopulasjonene, hvor de rømte laksene

kommer fra, må være svært forskjellige fra deres respektive avlspopulasjoner samtidig som de har bevart en betydelig synkronitet, om en skal fravike disse fortolkningene.

Resultatene så langt er derfor ikke i konflikt med den underliggende forklaringshypotesen for hvorfor det er så stor forskjell mellom antall rømt oppdrettslaks observert i elv og antall rømt oppdrettslaks som sådan, nemlig at oppdrettslaksen har blitt selektert for egenskaper som indirekte har forårsaket at en stor andel av rømt oppdrettslaks har en betydelig redusert ferskvannsoppvandringsevne.

4.1.4.4 Kobling mellom genetiske forskjeller og ferskvannsoppvandringsevne

Det tredje og siste resultatmålet for prosjektet var å avklare om de genetiske forskjellene **kan** kobles til biologiske mekanismer som underligger laksens evne (i vid betydning, se definisjon over) til å vende tilbake til ferskvann.

Prosedyre for identifisering av kandidat-SNPer: Dersom en kobling til slike biologiske mekanismer er tilstede, er det, som nevnt over, rimelig å anta at avlsarbeidet har selektert for genvarianter med lav frekvens i villaksen av negativ betydning for ferskvannsoppvandringsevne. Dette impliserer at de SNPene som eventuelt kan kobles til disse biologiske mekanismene (ved å være i sterk koblingsulikevekt med disse genene) er blant de hvor A1 er i lav til moderat frekvens hos villaks og hvor forskjellen i frekvens av A1 mellom villaks og rømt oppdrettslaks er liten. Det vil si alle rømte oppdrettslaks som greier å vandre opp i ferskvann er de som har genotyper og genotypekombinasjoner lik villaks for et spesifikt sett av SNPer.

Med utgangspunkt i dette filtrerte vi hele SNP-settet for de SNPene hvor forskjellen i A1-frekvens mellom rømtpopulasjon og villaks var mindre enn en gitt prosentverdi (EW), hvor A1-frekvensen i villaks var mindre enn en gitt prosentverdi (A1_max). På grunn av usikkerhet vedrørende underliggende genetisk arkitektur krevde vi, som en første tilnærming for å unngå muligheten for å dra inkonsistente slutninger, at SNPene var felles for alle rømtpopulasjonene. I og med at hvilke av disse felles-SNPene som har vært under direkte eller indirekte seleksjon kan være sterkt avlspopulasjonsavhengig, filtrerte vi deretter dette SNP-settet for SNPer hvor A1 hadde økt i frekvens i forhold til villaks over en gitt terskelverdi (AW) for hver de fire avlspopulasjonene.

Under følger resultatene for noen terskelverdikombinasjoner:

1. **For EW = 5, A1_max = 20 og AW = 15** for alle avlspopulasjonene identifiserte vi 27, 2, 11 og 2 SNPer for henholdsvis AquaGen (AQ), SalmoBreed (SB), Rauma (R) og Mowi (MH), med overlappingen:

   SNPs shared between AQ and SB: 0
   SNPs shared between AQ and R: 0
   SNPs shared between AQ and MH: 0
   SNPs shared between SB and R: 2
   SNPs shared between SB and MH: 2
   SNPs shared between R and MH: 2

2. **For EW = 5, A1_max = 50 og AW = 15** for alle avlspopulasjonene identifiserte vi 84, 5, 27 og 11 SNPer for henholdsvis AquaGen, SalmoBreed, Rauma og Mowi, med overlappingen:

   SNPs shared between AQ and SB: 0
   SNPs shared between AQ and R: 2
   SNPs shared between AQ and MH: 1
   SNPs shared between SB and R: 2
   SNPs shared between SB and MH: 2
   SNPs shared between R and MH: 2

3. **For EW = 10, A1_max = 20 og AW = 15** for alle avlspopulasjonene identifiserte vi 459, 14, 81 og 57 SNPer for henholdsvis AquaGen, SalmoBreed, Rauma og Mowi, med overlappingen:

   SNPs shared between AQ and SB: 1
   SNPs shared between AQ and R: 15
   SNPs shared between AQ and MH: 7
   SNPs shared between SB and R: 3
   SNPs shared between SB and MH: 7
   SNPs shared between R and MH: 12

4. **For EW = 10, A1_max = 30 og AW = 15** for alle avlspopulasjonene identifiserte vi 714, 25, 156 og 116 SNPer for henholdsvis AquaGen, SalmoBreed, Rauma og Mowi, med overlappingen:

   SNPs shared between AQ and SB: 2
   SNPs shared between AQ and R: 29
   SNPs shared between AQ and MH: 22
   SNPs shared between SB and R: 5
   SNPs shared between SB and MH: 8
   SNPs shared between R and MH: 21

5. **For EW = 10, A1_max = 50 og AW = 15** for alle avlspopulasjonene identifiserte vi 1351, 83, 372 og 313 SNPer for henholdsvis AquaGen, SalmoBreed, Rauma og Mowi, med overlappingen:

   SNPs shared between AQ and SB: 12
   SNPs shared between AQ and R: 73
   SNPs shared between AQ and MH: 64
   SNPs shared between SB and R: 8
   SNPs shared between SB and MH: 15
   SNPs shared between R and MH: 59

6. **For EW = 10, A1_max = 30** og **AW = 15** for avlspopulasjonene SalmoBreed, Rauma og Mowi og AW = 20 for AquaGen identifiserte vi 276, 25, 156 og 116 SNPer for henholdsvis AquaGen, SalmoBreed, Rauma og Mowi, med overlappingen:

   SNPs shared between AQ and SB: 1
   SNPs shared between AQ and R: 3
   SNPs shared between AQ and MH: 7
   SNPs shared between SB and R: 5
   SNPs shared between SB and MH: 8
   SNPs shared between R and MH: 21

Som vi ser er antallet SNP-kandidater som passerer disse filtrene svært følsomt for hvilke filtreringskriterier en bruker. En må derfor arbeide videre med SNP-sett som en etter beste skjønn mener har en tilstrekkelig størrelse samtidig som de ikke krever for store dataressurser å håndtere, og vi valgte sett nummer 3 som et første utgangspunkt.

Selv om vi i den første filtreringen krevde at SNPene skulle være felles for alle populasjonene, ser vi at kun en marginal andel SNPer er felles etter den andre filtreringen. Dette kan skyldes at avlsarbeidet med de ulike populasjonene har kapitalisert på ulike genvarianter, men disse ulike genvariantene kan i så fall likevel være koblet i den underliggende fysiologien.

Prosedyre for å knytte SNP-kandidater til biologiske mekanismer assosiert med ferskvannsoppvandrings-evne: For å kunne gjøre en slik kobling på troverdig vis, må en redusere størrelsen på kandidatsettene slik at en unngår å dra feilaktige konklusjoner basert på tilfeldigheter. En slik reduksjon vil måtte baseres på en konseptuell modell for hvordan tapet av ferskvannsoppvandringsevne har skjedd som følge av avlsarbeidet.

Filtreringsresultatene over er basert på en forestilling om at avlsarbeidet, enten ved direkte seleksjon eller indirekte på grunn av sterk koblingsulikevekt med genvarianter under direkte seleksjon, har økt frekvensen av genvarianter som i utgangspunktet er i lav frekvens i villakspopulasjonen og som er dysfunksjonelle for ferskvannsoppvandringsevnen. Resultatene over er i overenstemmelse med denne forestillingen, men de utelukker ikke andre scenarier.

Men uansett scenario, gitt at det virkelig er en kobling mellom biologiske mekanismer knyttet til ferskvannsoppvandringsevne og SNP-settene over, så må det finnes SNP-genotypekombinasjoner i dem som forefinnes i en langt høyere frekvens i villaks og rømt oppdrettslaks fra en gitt opphavspopulasjon enn opphavspopulasjonen, og som kan assosieres med gener av betydning for denne evnen.

Å bekrefte eller avkrefte prediksjonen over representerer en betydelig kombinatorisk utfordring. Vi valgte å ta utgangspunkt i at enhver SNP-genotypekombinasjon i et individ kan uttrykkes som en logisk struktur, det vil si enkeltgenotypene er bundet sammen med logiske operatorer i hvert individ til et logisk uttrykk, og at disse kan samles i et felles logisk uttrykk som favner alle individene i en populasjon.

Ut fra vurderingene over, og det faktum at ingen enkelt-SNPer kunne forklare dataene, valgte vi å systematisk filtrere et sett av kandidat-SNPer ved å starte med å bestemme dekningsgraden av alle 2-SNP-kombinasjoner av genotyper bundet sammen med den logiske operatoren AND i en gitt avlspopulasjon, i de rømte oppdrettslaksene tilordnet denne avlspopulasjonen og i villakspopulasjonen. Ut fra innledende studier forlangte vi at dekningsgraden for villaks og rømt oppdrettslaks skulle være minimum 50 % og dekningsgraden i avlspopulasjonen skulle være maksimalt 30 %. Deretter valgte vi ut de 2-SNP-genotypekombinasjonene som var høyest representert i villakspopulasjonen, samlet alle distinkte SNPer i et nytt sett hvorfra vi genererte alle mulige 3-SNP-genotypekombinasjoner bundet sammen med den logiske operatoren AND. I dette tilfellet forlangte vi at dekningsgraden for villaks og rømt oppdrettslaks skulle være minimum 25 % og dekningsgraden i avlspopulasjonen skulle være maksimalt 10 %. Disse terskelverdiene kan selvfølgelig justeres i senere og mer inngående analyser.

Deretter plukket vi de 3-SNP-genotypekombinasjonene med høyest dekningsgrad i villaks, og satte de sammen i alle mulige 2 x 3-SNP-kombinasjoner bundet sammen med den logiske operatoren OR. Fra disse kombinasjonene plukket vi igjen ut de med høyest dekningsgrad i villaks og satte sammen alle mulige (2- 4) x 2 x 3-SNP-genotypekombinasjoner, det vil si opptil 8 tretupler bundet sammen med OR og innbyrdes bundet sammen med AND. Ingen føringer på dekningsgraden i avlspopulasjonen ble gjort i de to siste trinnene. De opptil 8 3-tuppelkombinasjonene med høyest dekningsgrad i villakspopulasjonen og rømtlakspopulasjonen ble så slått sammen, og settet av distinkte SNPer fikk status som et førstegenerasjons SNP-sett som en kunne bruke for å teste om det var en troverdig kobling til biologiske mekanismer knyttet ferskvannsoppvandringsevne eller ikke. De høyeste dekningsgradene for 3-tuppelfiltreringen lå i området 60-80 % for villaks og rømtlaks, for kombinasjonene av 3-tupler lå de høyeste dekningsgradene på 100 % for villaks, 92- 100 % for rømtlaks og de assosierte dekningsgradene for avlspopulasjonen lå mellom 15-35 %.

Grunnen til å gjøre denne silingen av SNPer trinnvis var å unngå kombinatorisk eksplosjon og dermed svært lange beregningstider. En kan risikere å miste informasjon med denne tilnærmingen, men dette kan i ettertid kompenseres for ved å kjøre analysene på nytt og gjøre seg bruk av parallellprosessering for å redusere regnetiden.

Det er viktig å understreke at denne fremgangsmåten ikke impliserer et krav om at den virkelige logiske strukturen i datasettet består av slike tretupler av genotyper innbyrdes bundet sammen med en AND operator. Den impliserer heller ikke et krav om at en enkelt 3-SNP-genotype avgjør om et individ har bevart ferskvannsoppvandringsevnen eller ikke. Men den antar at kombinasjoner av minst tre SNP-genotyper vil i betydelig grad kunne skille avlspopulasjonsindivider fra villaks og rømt oppdrettslaks fanget i elv. Den

logiske modellen som ble brukt kan derfor betraktes som en førsteordens tilnærming for å avdekke om det virkelig finnes genotypekombinasjoner som har den karakteren en forventer ut fra den overordnede forklaringsmodellen som er motivasjonen for prosjektet.

Selv om vi legger spesifikke føringer i filtreringsprosedyrene over er denne fremgangsmåten legitim da disse føringene er en genuin test på om det virkelig eksisterer genotypekombinasjoner som skiller avlspopulasjonsindivider fra villaks og rømt oppdrettslaks fanget i elv. **Det er på ingen måte gitt på forhånd at vi skulle finne slike genotypekombinasjoner av et moderat antall SNPer som har tilnærmet 100 % dekningsgrad i villakspopulasjonen og rømtlakspopulasjonen og en lav dekningsgrad i avlspopulasjonen.**[1]

En ytterligere underbygging av dette er at vi i forbindelse med 3-tuppelfiltreringen kun satte som krav at dekningsgraden for villaks og rømt oppdrettslaks fanget i elv skulle være minimum 25 % og dekningsgraden i avlspopulasjonen skulle være maksimalt 10 % i de genotypekombinasjonene vi lot passere. Etterpå la vi ingen føringer. At vi får dekningsgrader på omtrent 100 % i villaks og rømtlaks og 30% i avlspopulasjon er derfor ikke et resultat av et forutinntatt søk for å bekrefte hypotesen vår. Slike dekningsgradsresultater følger heller ikke på noen måte direkte fra kravet om at A1-frekvensforskjellen mellom rømtlaks og villaks skulle være moderat lav for de SNPene vi lot passere i den aller første filtreringen, og at A1-frekvensen måtte ha økt moderat i den assosierte avlspopulasjonen i forhold til villaks.

Eksempel på annotering av et SNP-kandidatsett: Til eksemplifisering av utfallet av fremgangsmåten over valgte vi å bruke SNP-settet som ble produsert av parametersettet **EW = 10, A1_max = 20 og AW = 15 for alle fire populasjoner.** Antallet SNPer i de fire kandidatsettene fordelte seg slik: AquaGen: 18, Mowi: 16, SalmoBreed: 13, Rauma 13. Detaljer om disse SNPene er gitt i de fire listene under:

AquaGen:

	Cigene_id	Criteria	Chrom	Position	Annotation	Loc_info
16587	ctg7180001335371_3683_SAC	Distribution-SNP	ssa01	125843487	intergenic_region	LOC106565750-LOC106563392
34	ctg7180001885038_4967_SAC	rsb	ssa02	9269382	upstream_gene_variant	acbd7
5761	ctg7180001793773_4907	functional	ssa06	38936265	missense_variant	LOC106607255
23454	ctg7180001916146_1772_SCT	Distribution-SNP	ssa07	8952124	intron_variant	LOC106608572
25482	ctg7180001679281_5789_SAC	Distribution-SNP	ssa09	67730491	upstream_gene_variant	LOC106611890
6249	ctg7180001906456_6568_SGT	functional	ssa09	74670550	missense_variant	tmem129
6260	ctg7180001891409_3634_SCT	functional	ssa09	77924878	missense_variant	tmem173
276	ctg7180001805519_4864_SAG	rsb	ssa09	128631600	intron_variant	stim1
6570	ctg7180001870716_3401_SCT	functional	ssa10	74406853	missense_variant	saps3
6582	ctg7180001797454_43093_SAG	functional	ssa10	79356907	missense_variant	nup93
29040	ctg7180001845082_7172_SAG	Distribution-SNP	ssa11	47799462	intergenic_region	LOC106563007-LOC106562883
29727	ctg7180001281527_848_SCT	Distribution-SNP	ssa11	91377549	intron_variant	LOC106563916
2678	ctg7180001845450_3014_SAC	env,wild	ssa14	54777236	missense_variant	LOC106570060
7509	ctg7180001900376_8301_SAG	functional	ssa14	61498291	missense_variant	LOC106569887
7909	ctg7180001835143_8201_SGT	functional	ssa16	53681459	missense_variant	LOC106574182
470	ctg7180001912369_4971_SAG	rsb	ssa17	6071248	synonymous_variant	LOC106575232
8196	ctg7180001398466_7983_SAG	functional	ssa18	21420780	missense_variant	LOC106577055
9459	ctg7180001857660_10743_SAG	functional	ssa27	4563122	missense_variant	LOC106588229

[1] Referansegruppen har påpekt at bare man tester mange nok kombinasjoner så vil det alltid finnes genotypekombinasjoner som skiller en gruppe fisk fra en annen. Men i dette tilfellet tester vi positivt en spesifikk hypotese ved å sammenligne tre grupper fisk med hverandre i fire uavhengige populasjonsgrupper, noe som reduserer denne muligheten betydelig. Dessuten var det ikke mulig å finne det inverse mønsteret, det vil si en stor gruppe av genotypekombinasjoner hvor dekningsgraden var høy hos villaks og avlspopulasjon, og ikke hos rømtlaks fanget i elv. Men uansett er en slik bekreftelse kun legitimerende for en mer inngående biologisk og genetisk analyse.

SalmoBreed:

	Cigene_id	Criteria	Chrom	Position	Annotation	Loc_info
18035	ctg7180001853207_6271_SAG	Distribution-SNP	ssa02	65620891	intron_variant	LOC106593154
18970	ctg7180001631827_192_SAC	Distribution-SNP	ssa03	56012380	intron_variant	gbgt1
24625	ctg7180001903360_761_SAC	Distribution-SNP	ssa08	26237329	intron_variant	melk
25311	ctg7180001851039_10290_SAC	Distribution-SNP	ssa09	55068124	intergenic_region	LOC106611576-LOC106611580
6326	ctg7180001878854_3628_SAG	functional	ssa09	102397307	missense_variant	LOC106612683
6396	ctg7180001668257_1374_SAG	functional	ssa09	127219319	missense_variant	LOC106613166
32574	ctg7180001827868_2803_SAC	Distribution-SNP	ssa13	90658249	intron_variant	LOC106568138
7431	ctg7180001591304_4622_SGT	functional	ssa14	34030280	missense_variant	LOC106569383
33706	ctg7180001937818_13928_SAG	Distribution-SNP	ssa14	59744977	intron_variant	LOC106569936
34861	ctg7180001827628_2639_SGT	Distribution-SNP	ssa15	38579942	intron_variant	LOC106571504
34972	ctg7180001208887_1776_SCT	Distribution-SNP	ssa15	46306150	intron_variant	LOC106571584
41794	ctg7180001364094_5295_SCT	Distribution-SNP	ssa21	2283284	intergenic_region	dach1-gpr18
43412	ctg7180001474449_216_SGT	Distribution-SNP	ssa22	48503098	upstream_gene_variant	LOC106583629

Rauma:

	Cigene_id	Criteria	Chrom	Position	Annotation	Loc_info
21377	ctg7180001883233_10351	Distribution-SNP	ssa05	41056260	synonymous_variant	LOC106604953
23606	ctg7180001289769_2604_SCT	Distribution-SNP	ssa07	19281860	intron_variant	LOC106608718
23766	ctg7180001858455_7377_SAG	Distribution-SNP	ssa07	30522174	intron_variant	LOC100306760
10958	ctg7180001787680_1094_SAC	WildSouth-Aqua	ssa09	17186275	downstream_gene_variant	LOC106610816
10959	ctg7180001628780_3092_SAG	WildSouth-Aqua	ssa09	17216431	intergenic_region	LOC106610822-LOC106610817
26059	ctg7180001552685_1089_SGT	Distribution-SNP	ssa09	108793201	intron_variant	ift80
27707	ctg7180001934773_3779_SCT	Distribution-SNP	ssa10	76986708	intron_variant	zfpm1
7431	ctg7180001591304_4622_SGT	functional	ssa14	34030280	missense_variant	LOC106569383
37673	ctg7180001936744_2288_SCT	Distribution-SNP	ssa17	26667003	intergenic_region	LOC106575824-LOC106575825
37679	ctg7180001825403_10894_SAG	Distribution-SNP	ssa17	27027265	intergenic_region	LOC106575831-LOC106575832
37722	ctg7180001900396_1652_SCT	Distribution-SNP	ssa17	29730369	intron_variant	LOC106575937
40474	ctg7180001850393_8798_SCT	Distribution-SNP	ssa20	403615	upstream_gene_variant	LOC106579799
44228	ctg7180001809345_1074_SCT	Distribution-SNP	ssa23	36182565	upstream_gene_variant	LOC106584531

Mowi:

	Cigene_id	Criteria	Chrom	Position	Annotation	Loc_info
17077	ctg7180001933798_3460_SGT	Distribution-SNP	ssa01	158726051	intergenic_region	trnaa-ugc-LOC106572858
17230	ctg7180001888084_235_SCT	Distribution-SNP	ssa02	9446772	upstream_gene_variant	LOC106576116
17249	ctg7180001295039_234_SAG	Distribution-SNP	ssa02	10562168	intron_variant	LOC106576689
17337	ctg7180001593480_3370_SAC	Distribution-SNP	ssa02	16771122	intergenic_region	LOC106579240-LOC106579246
18970	ctg7180001631827_192_SAC	Distribution-SNP	ssa03	56012380	intron_variant	gbgt1
25311	ctg7180001851039_10290_SAC	Distribution-SNP	ssa09	55068124	intergenic_region	LOC106611576-LOC106611580
28165	ctg7180001626552_1091_SAG	Distribution-SNP	ssa10	107341104	intron_variant	LOC106561641
7431	ctg7180001591304_4622_SGT	functional	ssa14	34030280	missense_variant	LOC106569383
36633	ctg7180001833869_486_SCT	Distribution-SNP	ssa16	47245536	intergenic_region	LOC106574075-LOC106574044
42679	ctg7180001925561_4576_SAC	Distribution-SNP	ssa22	1397901	intron_variant	LOC106582631
44199	ctg7180001538584_4406_SGT	Distribution-SNP	ssa23	34429256	intergenic_region	LOC106584451-LOC106584480
9051	ctg7180001863290_453_SCT	functional	ssa23	35950060	missense_variant	mettl20
44228	ctg7180001809345_1074_SCT	Distribution-SNP	ssa23	36182565	upstream_gene_variant	LOC106584531
46880	ctg7180001268296_4006_SCT	Distribution-SNP	ssa27	23010617	intron_variant	LOC106588773
47482	ctg7180001903803_1531_SCT	Distribution-SNP	ssa28	18550835	intron_variant	cdrt1
47672	ctg7180001721295_376_SCT	Distribution-SNP	ssa28	30107494	intergenic_region	LOC106589868-LOC106589902

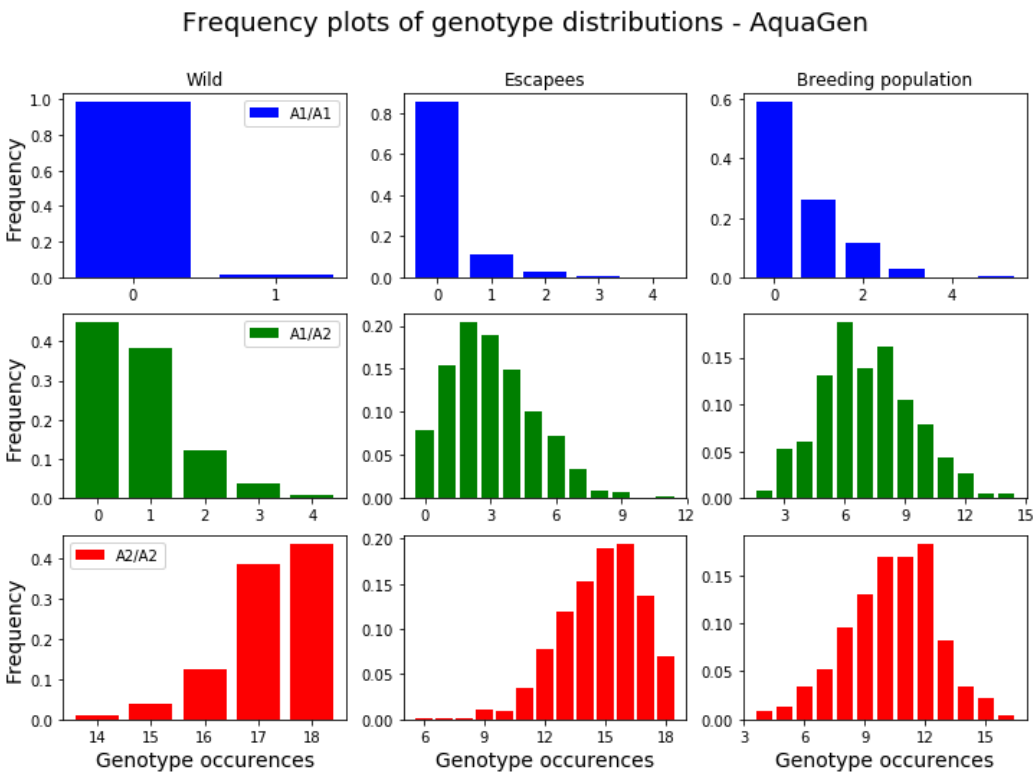
Selv om en kandidat-SNP er i ett gen, er det ingen garanti for at dette genet er assosiert med ferskvannsoppvandringsevne. En er derfor avhengig av å lete etter gener som ligger nær SNPen, og derved i sterk koblingsulikevekt med den, for å sjekke ut hvilken av dem, om noen, er assosiert med relevante biologiske mekanismer. Vi valgte å bruke et vindu på ±50 kilobaser (Kb). I en senere oppfølging kan en øke dette vinduet for å være enda sikrere på å få med seg alle relevante gener, men erfaring tilsier at i de fleste tilfeller vil en få med seg det meste med ±50 Kb.

Resultatene fra denne systematiske søkingen for de fire SNP-settene over er gitt i Vedlegg 3. Hovedkonklusjonen er at de genetiske forskjellene mellom rømt oppdrettslaks fanget i elv og avlspopulasjonene de kommer fra kan kobles til biologiske mekanismer som påvirker laksens evne til å vende tilbake til ferskvann som sådan (luktbiologi, utvikling av nervesystem, signalprosessering i hjernen, hukommelse, læring, etc). Men flere av de identifiserte genene kan også ha en direkte effekt på evne til å overleve i sjøvannsfasen, slik som å unngå predasjon eller evne til å fange byttedyr. Annoteringsresultatene indikerer at en og samme genotype kan i flere tilfeller påvirke begge disse to egenskapsklassene samtidig, noe som gjør det svært utfordrende å genetisk skille disse to klassene fra hverandre.

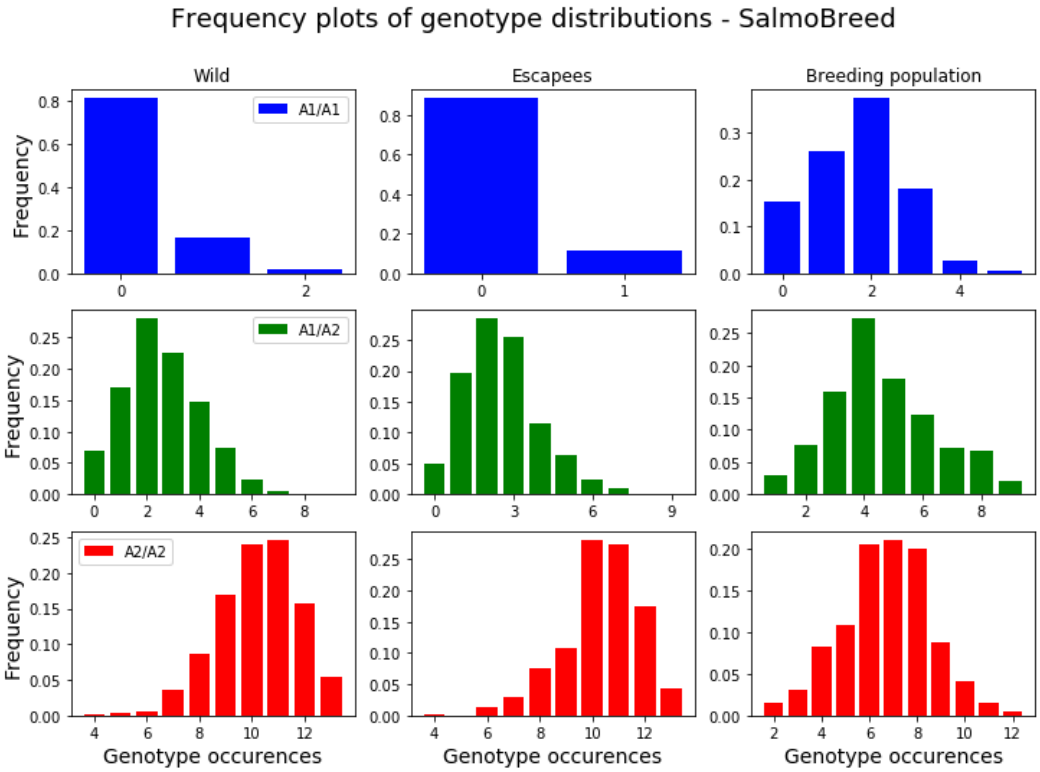
Selv om søket etter koblinger ble forsøkt utført så etterrettelig som mulig kan likevel ikke disse resultatene brukes til å hevde at SNPene virkelig **er** koblet til biologiske mekanismer assosiert med ferskvanns-oppvandringsevne. Men en tilnærming som vil bringe oss betydelig nærmere en slik avklaring er at en gjør et systematisk søk som over så godt som overhodet mulig, og ut fra resultatene formulerer spesifikke hypoteser om i hvilke gener en forventer å finne funksjonell genetisk variasjon som markant skiller avlspopulasjonene fra rømt oppdrettslaks fanget i elv og villaks. Om en deretter ved resekvensering av de identifiserte kandidatgenene finner SNP-varianter med høyst sannsynlig negativ funksjonell effekt på ferskvannsoppvandringsevne, og at disse er tilstede i langt høyere frekvens i avlspopulasjonene enn i rømtlaks og villaks, er det legitimt å konkludere at disse genene **kan være** av kausal betydning for ferskvannsoppvandringsevne.

Men selv positive resekvenseringsresultater vil ikke gi noe informasjon om i hvilken genotypekontekst en potensielt kausal genvariant må befinne seg i for å bidra til hemming av ferskvannsoppvandringsevne. Selv om det ligger utenfor prosjektets resultatmål å avdekke slike genotype-kombinasjoner, har vi utfra hensyn til praktisk nytteverdi gjort et omfattende programmeringsarbeid for å avdekke hvilke kombinasjoner som vi, selv uten tilgang til resekvenseringsinformasjon, kan tillegge betydelig kausal vekt. Vi har valgt å ikke presentere alle resultatene fra dette arbeidet før de er noe mer gjennomarbeidet. I det følgende presenterer vi derfor kun noen innledende resultater, men som likevel sannsynliggjør at det kan være mulig å anvende den frembrakte kunnskapen (kombinert med resekvenseringsinformasjon) i genombasert presisjonsavl.

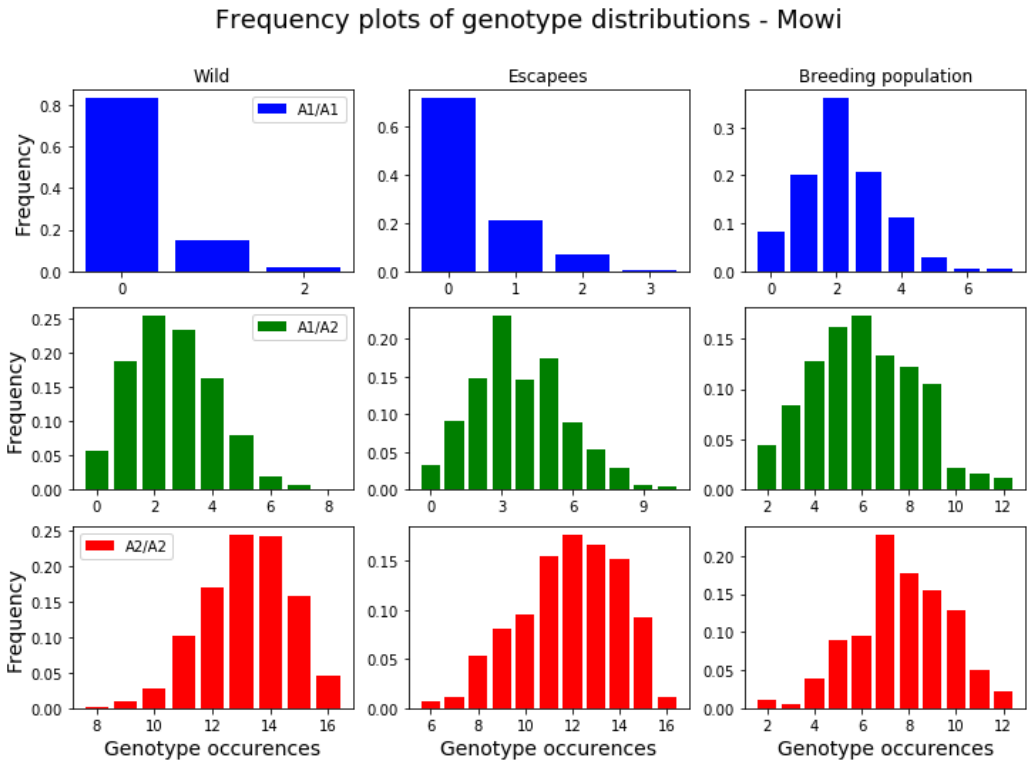
Genotypedistribusjoner: Et naturlig startpunkt for å tilnærme seg kausalitetsvurderinger er å få en oversikt over genotypedistribusjonene i de populasjonsgruppene de fire SNP-settene over er assosiert med. Genotypedistribusjonen for de fire populasjonsgruppene er gitt i Figur 11, 12, 13 og 14. Vi ser at for samtlige grupper så ligger de rømte oppdrettslaksene (escapees) nærmere villaksen enn opphavspopulasjonen.



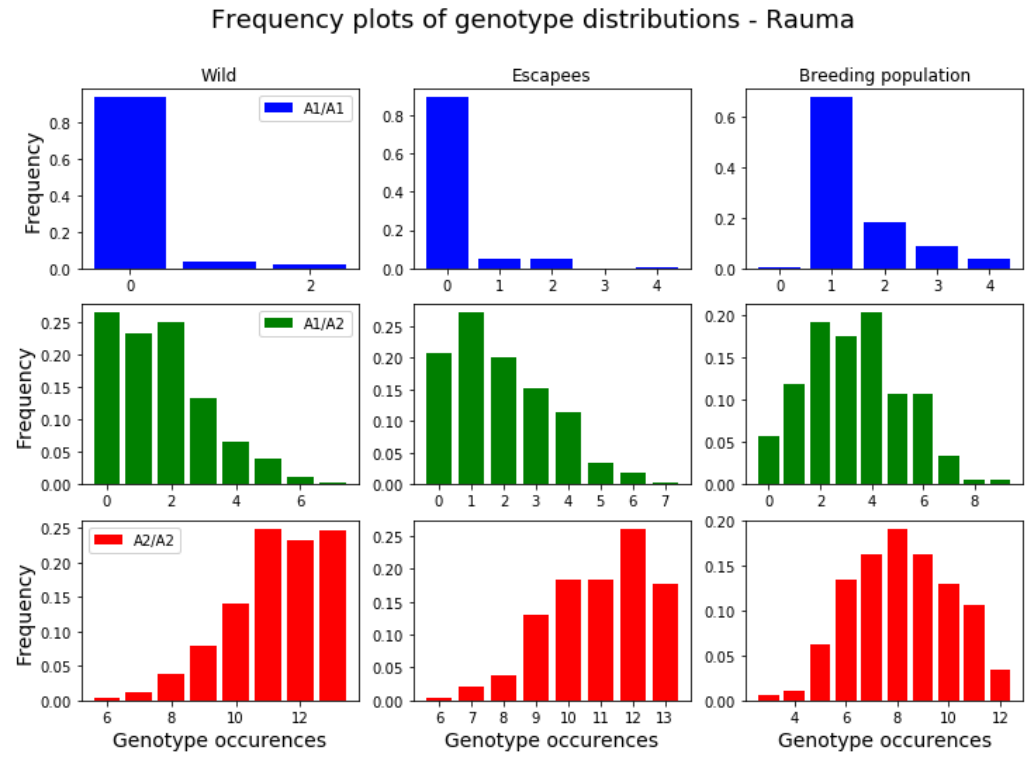
Figur 11: Genotypedistribusjonen for AquaGen-gruppen



Figur 12: Genotypedistribusjonen for SalmoBreed-gruppen

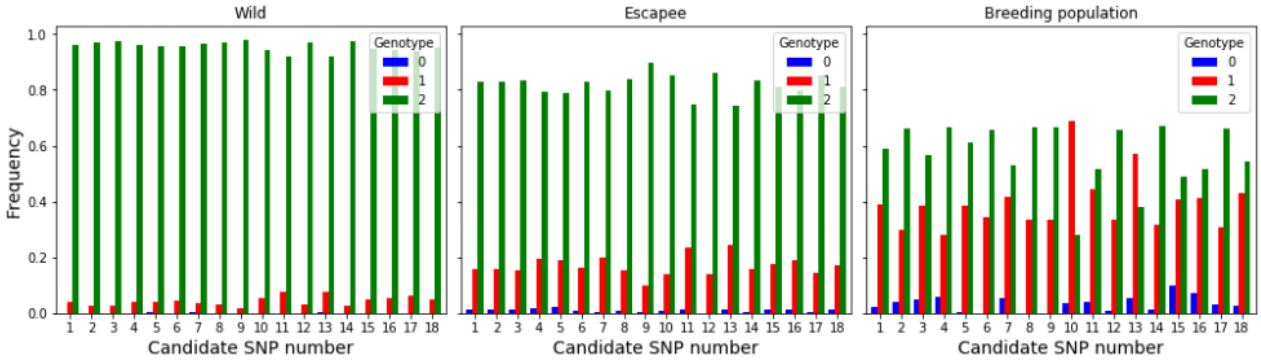


Figur 13: Genotypedistribusjonen for Mowi-gruppen

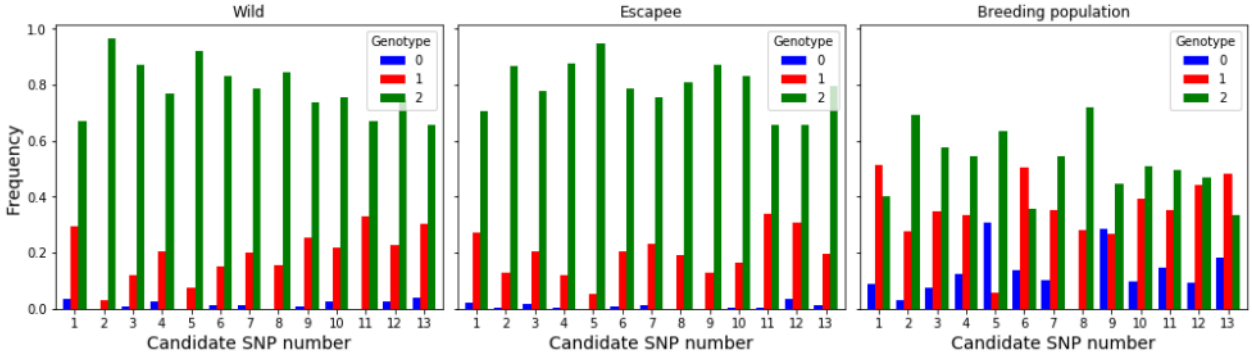


Figur 14: Genotypedistribusjonen for Rauma-gruppen

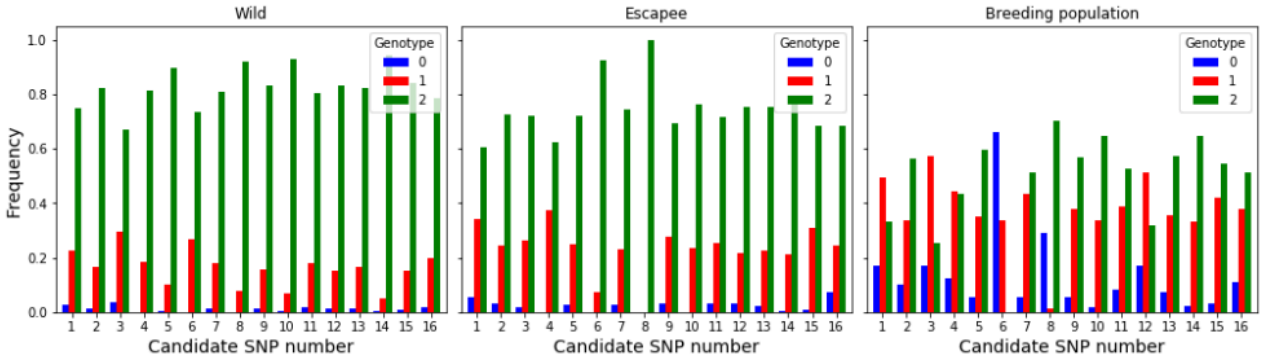
Frekvensfordelingen til de ulike genotypene for hver SNP i kandidatsettet for de tre ulike populasjonene i hver populasjonsgruppe gir ytterligere innsikt (Figur 15, 16, 17 og 18).



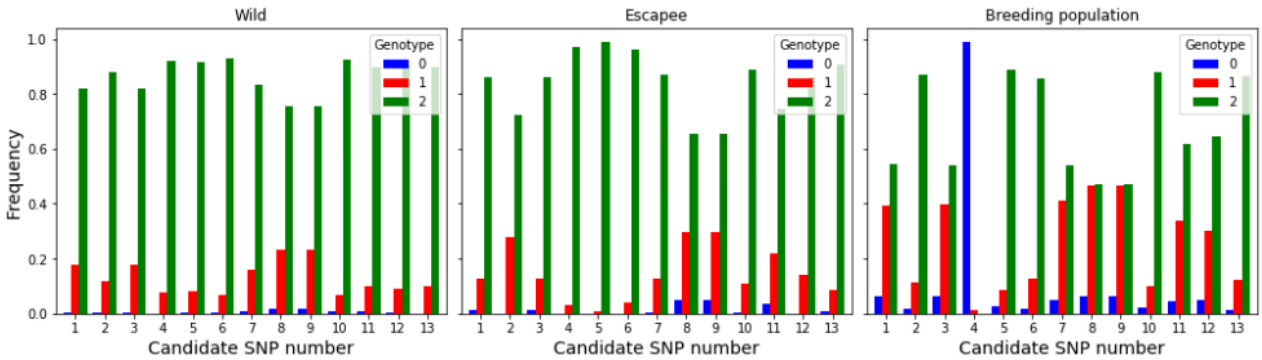
Figur 15: SNP-vis genotypefordeling i AquaGen.



Figur 16: SNP-vis genotypefordeling i SalmoBreed



Figur 17: SNP-vis genotypefordeling i Mowi.



Figur 18: SNP-vis genotypefordeling i Rauma.

Vi ser at de fire siste plottene viser med tydelighet hvor mer lik rømt oppdrettslaks fanget i elv er villaks enn opphavspopulasjonen sin for det gitte SNP-settet. Andelen heterozygoter og homozygoter for A1 er gjennomgående høyere i avlspopulasjonene enn i de to andre populasjonene. Rauma skiller seg dog noe ut, og andelen A1A1 homozygoter i SNP4 er påfallende høy.

Genotypeunikhetsgraden i datasettene: En naturlig oppfølging er å spørre om hvor mange distinkte genotyper for det aktuelle SNP-settet det er i de ulike populasjonsgruppene (SNP-settene er listet på side 23-24). Prosentandelen duplikater i henholdsvis villaks, rømt oppdrettslaks og avlspopulasjon fordeler seg slik for de fire gruppene for sine respektive SNP-sett:

AquaGen: 84.14, 28.54, 4.78
SalmoBreed: 45.34, 43.54, 0.51
Mowi: 38.16, 13.78, 0.55
Rauma: 79.38, 72.70, 41.01

Vi ser at det er en betydelig variasjon. I alle tilfellene er avlspopulasjonen langt mer heterogen enn de to andre gruppene. AquaGen og Rauma utmerker seg ved at villakspopulasjonen for de to aktuelle SNP-settene har usedvanlig mange felles genotyper. Rauma skiller seg også ut ved at den rømte oppdrettslaksen har en homogenitet på linje med villaksen og at avlspopulasjonen også har en betydelig homogenitet.

Boolsk beskrivelse av genotypestrukturen: Over filtrerte vi populasjonene basert på OR-kombinasjoner av tretupler innbyrdes bundet sammen med en AND-operator. Som tidligere fremhevet er et slikt Boolsk uttrykk kun å betrakte som en approksimering av den underliggende logiske strukturen i SNP-genotypesettene. Den virkelige strukturen kan finnes med teoretiske verktøy primært utviklet for konstruksjon av mikroprosessorer. Men på grunn av at vi har tre forskjellige genotypeverdier (0, 1 og 2 for å beskrive A1A1, A1A2 og A2A2), så kan ikke genotypematrisen for en gitt populasjon binariseres uten å miste informasjon. Skal en unngå dette må en gjøre seg nytte av multiverdilogikk som kan håndtere mer enn to verdier for en inputvariabel. Det finnes avansert teori for slik logikk, men vi har ikke greid å finne tilgjengelig programvare som kan gjøre slike analyser, og det er ytterst krevende å lage algoritmer for slik logisk analyse fra bunnen av. Vi fokuserte derfor i første omgang på å gjøre bruk av binariserte genotypematriser for å se hvor mye innsikt dette kunne gi.

Vi startet ut med å binarisere slik at alle A2A2-genotyper for ett individ fikk verdien 1, mens heterozygoten og den andre homozygoten fikk verdien 0. Så bestemte vi det Boolske utrykket som beskrev alle individene i rømtpopulasjonen for hver gruppe. Her er ett utdrag fra et slik uttrykk (som er betydelig lengre):

(SNP1 & SNP10 & SNP13 & SNP2 & SNP3 & SNP4 & SNP5 & SNP6 & SNP9) | (SNP10 & SNP11 & SNP13 & SNP2 & SNP3 & SNP4 & SNP5 & SNP6 & SNP9) | (SNP11 & SNP2 & SNP3 & SNP4 & SNP5 & SNP6 & SNP7 & SNP8 & SNP9) | (SNP1 & SNP10 & SNP11 & SNP12 & SNP13 & SNP5 & SNP6 & SNP7 & SNP8 & SNP9) | (SNP1 & SNP10 & SNP11 & SNP12 & SNP2 & SNP3 & SNP4 & SNP5 & SNP8 & SNP9) | (SNP1 & SNP10 & SNP11 & SNP12 & SNP2 & SNP4 & SNP5 & SNP6 & SNP7 & SNP8) | (SNP1 & SNP10 & SNP11 & SNP12 & SNP2 & SNP4 & SNP5 & SNP7 & SNP8 & SNP9) | ……

Det finnes sofistikerte, om enn ytterst beregningskrevende, metoder for å minimere slike uttrykk, men vi har ikke hatt tilgang på nok datakraft til å gjøre dette, og denne øvelsen er uansett ikke sentral for vårt bruk.

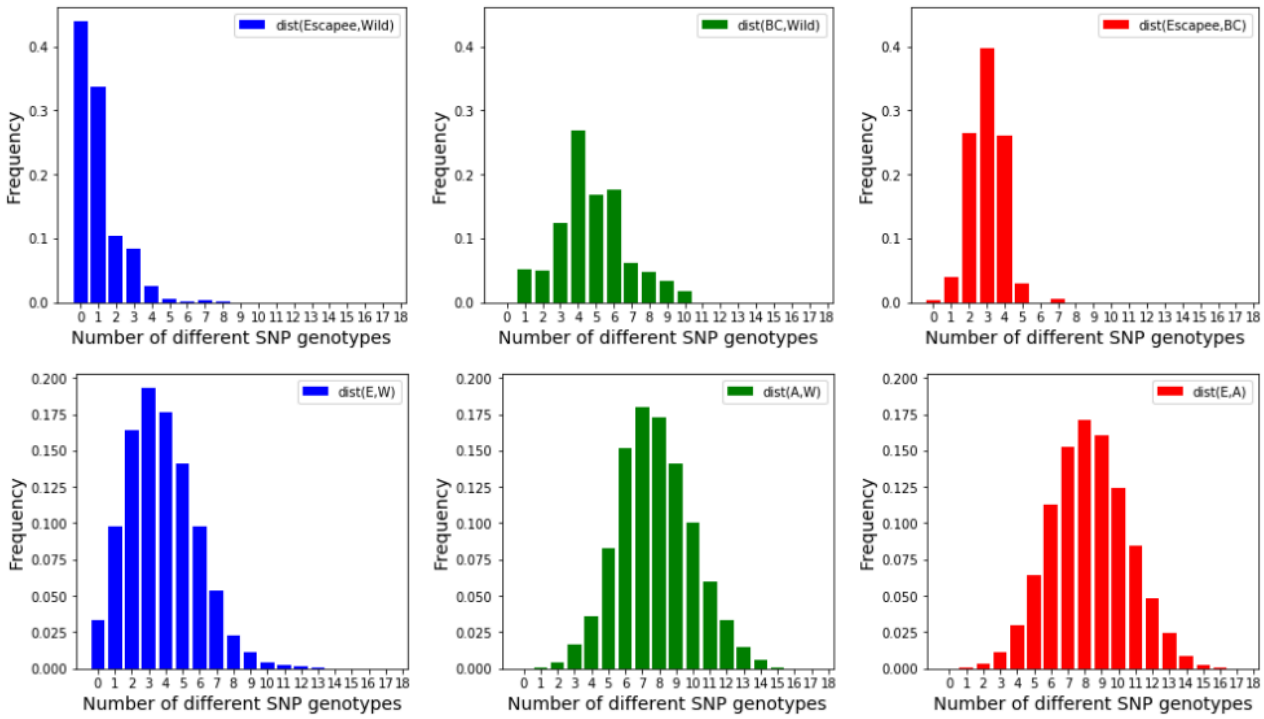
Deretter gjorde vi om det Boolske uttrykket for hver rømtlaksgruppe til en funksjon som vi kunne bruke til å teste hvor stor prosentandel av villaksindividene og avlspopulasjonsindividene ga verdien Sann når vi ga funksjonen de binariserte genotypeverdiene til individer fra disse to gruppene. Vi fant følgende:

AquaGen rømt mot avlspopulasjon: 2.61 %, AquaGen rømt mot villaks: 84.46 %

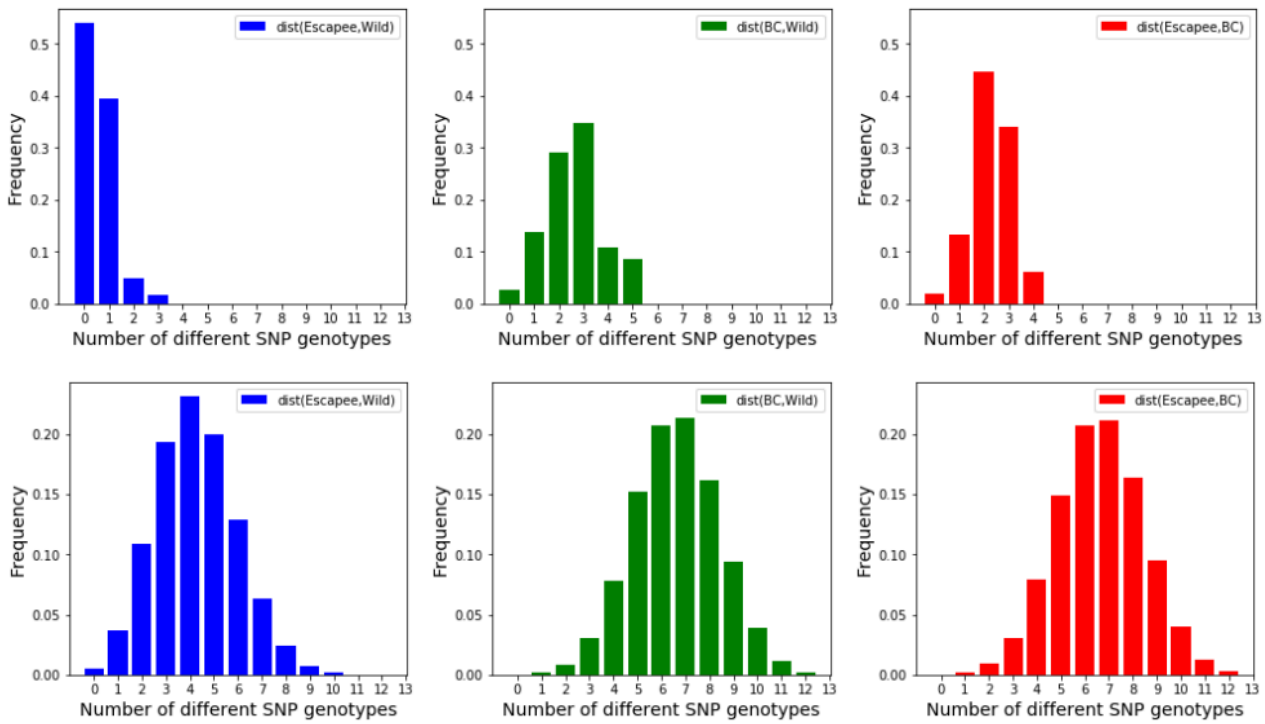
SalmoBreed rømt mot avlspopulasjon:	3.59 %,	SalmoBreed rømt mot villaks:	55.71 %
Mowi rømt mot avlspopulasjon:	0.0 %,	Mowi rømt mot villaks:	27.17 %
Rauma rømt mot avlspopulasjon:	10.11 %,	Rauma rømt mot villaks:	81.08 %

Vi ser at det er en stor ulikhet i den logiske strukturen mellom rømt oppdrettslaks fanget i elv og avlspopulasjon, og langt større likhet med villaks. Men Mowi skiller seg kraftig ut ved at det kun er 27 % av villaks for det gitte SNP-settet som har en match mot rømtlaksstrukturen. Resultatet er dog i overensstemmelse med genotypehomogenitetstallene over. Den reduksjonen av oppløsningen til genotypeinformasjonen som vi måtte gjøre, vil nødvendigvis føre til lavere dekningsgrad enn den sanne verdien. I tillegg vil tilfeller hvor den rømte oppdrettslaksen har genotypeverdien A2A2 og villaksen verdien A1A2 ikke gi match selv om de to genotypene sannsynligvis har samme fenotype. Det er derfor behov for å utvide analysen med å se nøyere på genotypesammensetningen som sådan mellom populasjonene i hver populasjonsgruppe for å kunne ta høyde for dominanseffekter og interaksjonseffekter, og så gjøre seg bruk av et nytt lag med logiske analyser ut fra dette.

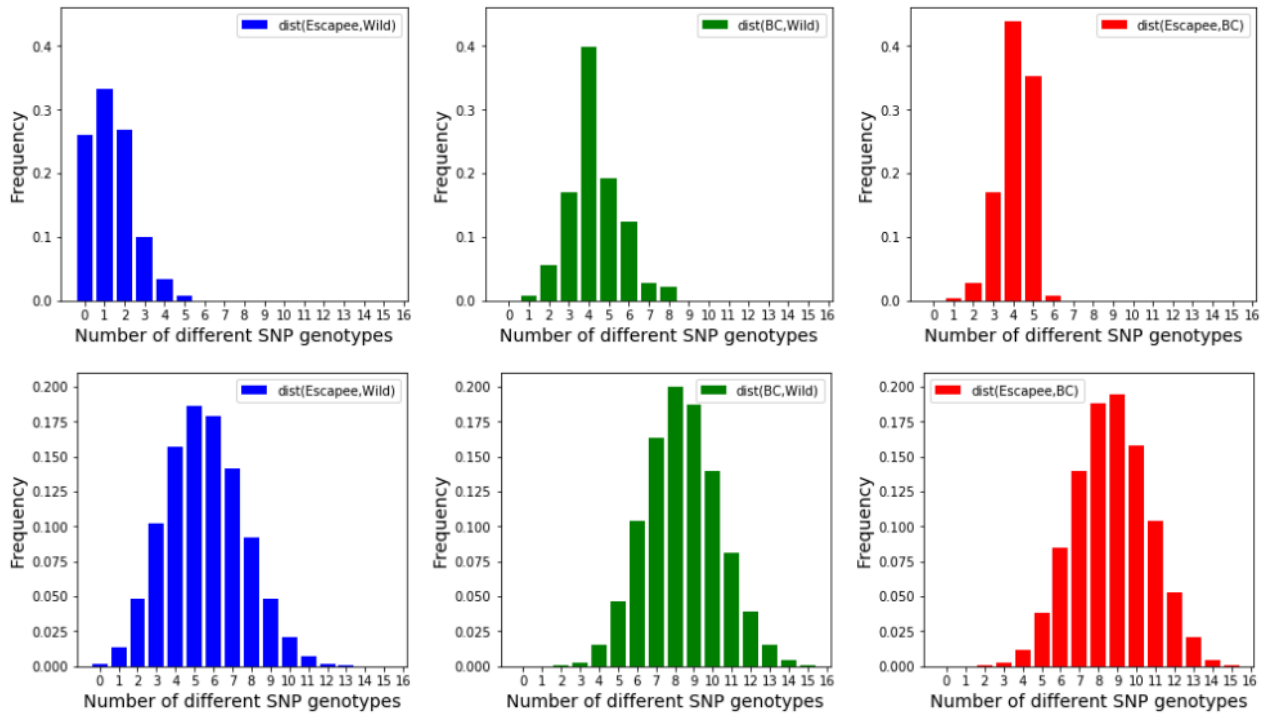
Levenshteinmetrikk-analyser: En måte å øke oppløsningen på for å forberede dypere logiske analyser er å gjøre seg nytte av såkalt Levenshteinmetrikk. Levenshtein-distansen er en strengmetrikk for å beskrive forskjellen mellom to sekvenser av tegn, det vil si minimum antall ett-tegns redigeringer (i vårt tilfelle kun substitusjoner) som trengs for å forandre en sekvens til å bli lik den andre. Den kan brukes til å vise hvor forskjellig genotypestrukturen er mellom populasjoner. Det øverste panelet i Figur 19, 20, 21 og 22 viser for hver populasjonsgruppe frekvensen av Levenshtein-distanser når en for hvert individ i rømtlakspopulasjonen måler distansen til nærmeste individ i villakspopulasjonen (venstre panel), for hvert individ i avlspopulasjonen måler distansen til nærmeste individ i villakspopulasjonen (midtre panel), og for hvert individ i rømtlakspopulasjonen måler distansen til nærmeste individ i villakspopulasjonen (høyre panel). Det nederste panelet i Figur 19, 20, 21 og 22 viser for hver populasjonsgruppe frekvensen av Levenshtein-distanser når en for hvert individ i første populasjon (dist(pop1,pop2)) måler distansen til alle individer i andre populasjon.



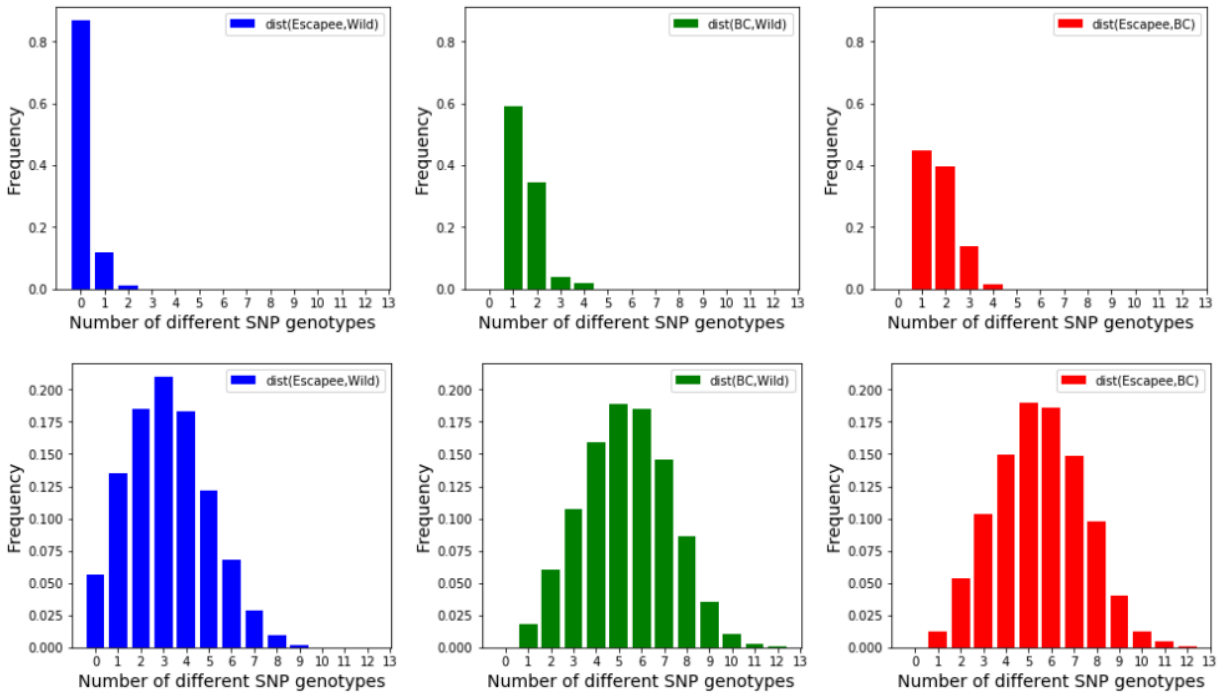
Figur 19: Frekvensfordelinger av Levenshtein-distansen mellom rømt oppdrettslaks og villaks, mellom avlspopulasjon og villaks, og mellom rømt oppdrettslaks og avlspopulasjon for AquaGen. Se tekst for hva som skiller øverste og nederste panel. SNP-settet er listet på side 23.



Figur 20: Frekvensfordelinger av Levenshtein-distansen mellom rømt oppdrettslaks og villaks, mellom avlspopulasjon og villaks, og mellom rømt oppdrettslaks og avlspopulasjon for SalmoBreed. SNP-settet er listet på side 24.



Figur 21: Frekvensfordelinger av Levenshtein-distansen mellom rømt oppdrettslaks og villaks, mellom avlspopulasjon og villaks, og mellom rømt oppdrettslaks og avlspopulasjon for Mowi. SNP-settet er listet på side 24.



Figur 22: Frekvensfordelinger av Levenshtein-distansen mellom rømt oppdrettslaks og villaks, mellom avlspopulasjon og villaks, og mellom rømt oppdrettslaks og avlspopulasjon for Rauma. SNP-settet er listet på side 24.

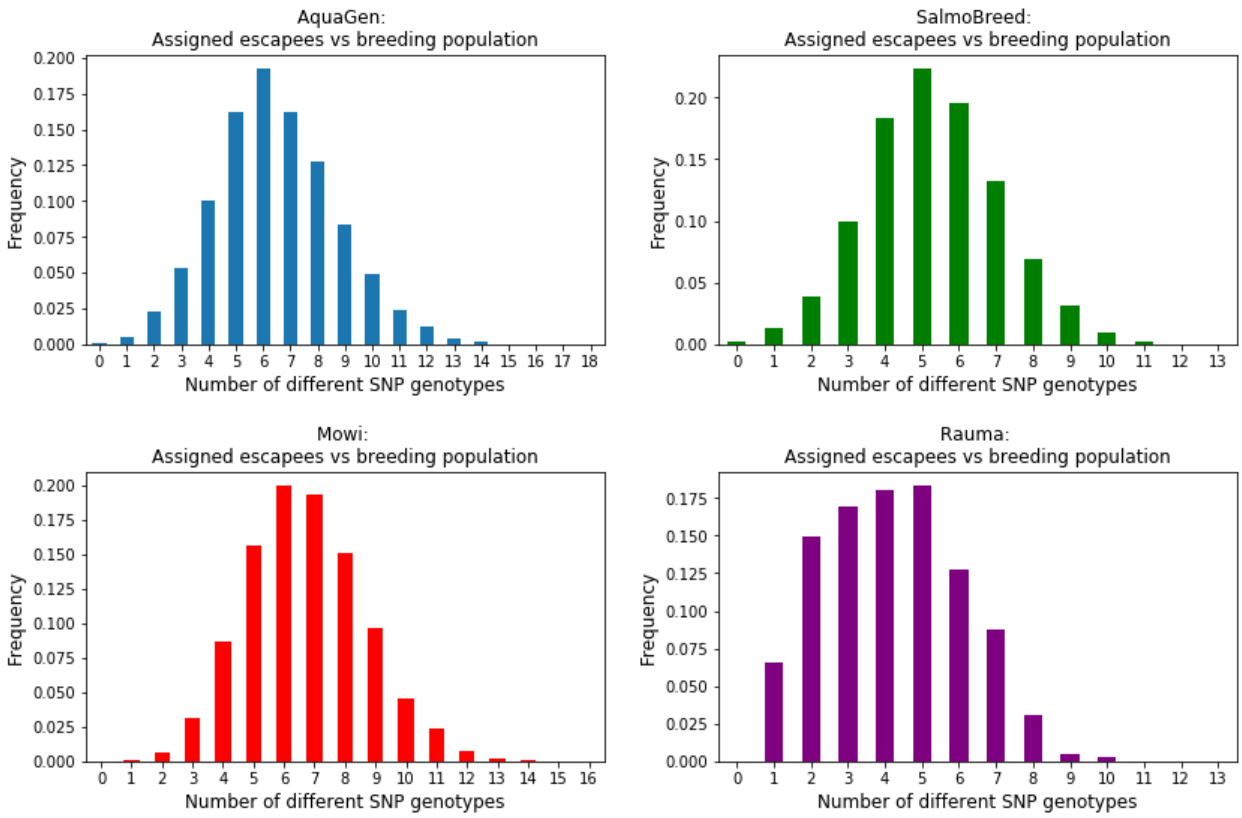
Plottene bekrefter resultatene over, men i tillegg gir de ytterligere informasjon om hvor store ulikheter det egentlig er i genotypestrukturen mellom de enkelte populasjonene, og de gir oss derfor ytterligere grunnlag for kausalitetsvurderinger. Denne tilnærmingen åpner derfor muligheter for å finne mønstre som er av betydning for nødvendighets- og tilstrekkelighets-vurderinger vedrørende SNP-genotypekombinasjoner som gir henholdsvis opprettholdelse og ikke opprettholdelse av ferskvannsoppvandringsevne.

I de tilfeller hvor SNP-genotypen til en rømt oppdrettslaks er A1A1 eller A1A2 og et individ fra avlspopulasjonen har genotypen A2A2, er det grunn til å hevde at dette locuset ikke har negativ effekt på ferskvannsoppvandringsevnen i avlspopulasjonsindividet. Når vi justerer distansemålene over for dette får vi en synlig endring i frekvensen av Levenshtein-distanser (Figur 23).

I Figur 23 er hvert individ i rømtpopulasjonen sammenlignet med alle individer i avlspopulasjonen. Ved at plottene maskerer forskjeller som med stor sannsynlighet ikke er av kausal betydning, er disse fordelingene etter vår mening langt mer relevante å analysere.

Oppvandringstall sammenlignet med rømningstall koblet med analyseresultatene så langt underbygger forestillingen om at en betydelig andel av individer fra avlspopulasjonene ikke er i stand til å komme tilbake til ferskvann om de rømmer. Fortolker vi Figur 23 utfra denne forestillingen så er det legitimt å hevde at desto flere SNPer som skiller et rømt individ som har vandret opp i elv fra et avlspopulasjonsindivid, desto større er sannsynligheten for at avlspopulasjonsindividet har en redusert ferskvannsoppvandringsevne. Dette utelukker dog ikke at det finnes eksempler på at kun en 2-SNP-forskjell vil være avgjørende. Men om vi bruker denne sannsynlighetsbetraktningen, og for eksempel antar at en 5-SNP-forskjell i snitt er **nødvendig** og **tilstrekkelig** for å fjerne ferskvannsoppvandringsevnen, så gir figuren oss et grovt estimat for andelene av avlspopulasjonene som har henholdsvis intakt og ikke intakt ferskvannsoppvandringsevne. Vi ser at bortsett fra Rauma er de tre fordelingene temmelig like, og at ferskvannsoppvandringsevnen til

Raumapopulasjonen var langt mer intakt i innsamlingsåret enn hos de tre andre utfra de foregående premissene.



Figur 23: Frekvensfordelingen av Levenshtein-distansen mellom rømt oppdrettslaks og avlspopulasjon etter å ha justert for ulikheter som antas å ikke ha negativ betydning for ferskvannsoppvandringsevne i avlspopulasjonsindivider. De brukte SNP-settene er listet på side 23-24.

Følgende **forklaringsmodell til underliggende genetisk arkitektur** er konsistent med funnene over:

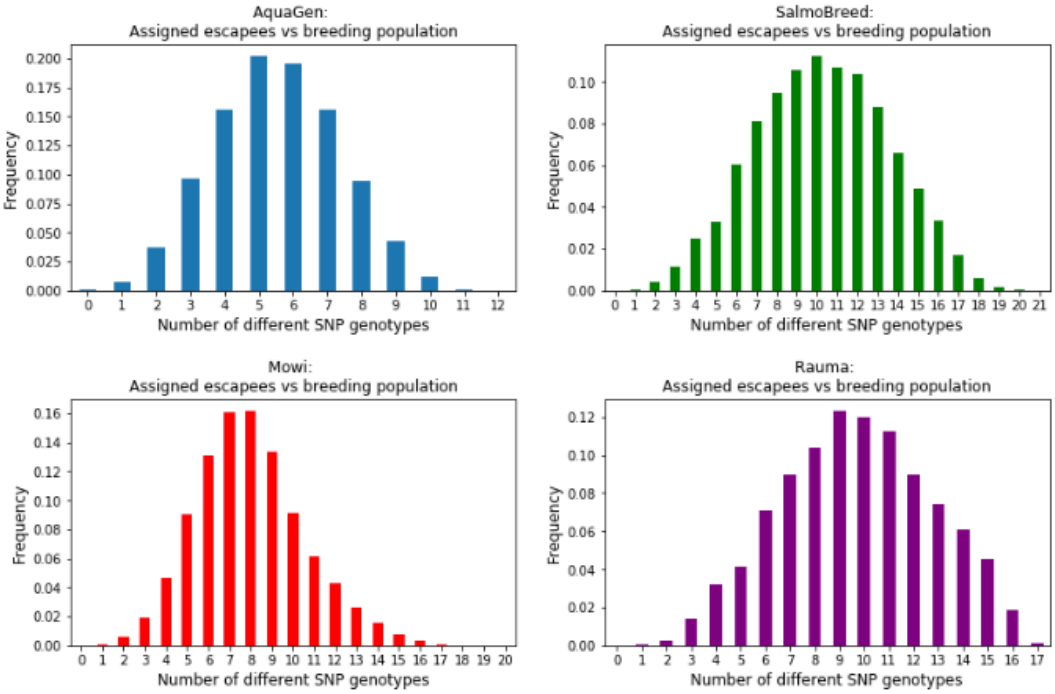
Avlsarbeidet har økt frekvensen av alleler (genvarianter) som forefinnes i lav frekvens i villakspopulasjonen. Dette kan skyldes både direkte og indirekte seleksjon, og til en viss grad også tilfeldig genetisk drift. Mange av disse allelene kan bidra til tap av ferskvannsoppvandringsevne, og det er derfor en polygen basis for denne egenskapen. Men kun et fåtall av disse allelene trenger å være samtidig tilstede i et individ for at dets ferskvannsoppvandringsevne blir redusert i svært stor grad. Egenskapen som sådan er derfor kausalt sett oligogen, selv om den har en polygen basis hvor de oligogene kombinasjonene trekkes fra. Genvariantene som inngår i en oligogen kombinasjon kan virke både additivt og ikke-additivt.[2]

Av de 48075 SNPene vi brukte i denne studien, er andel SNPer med en A1-frekvens på 5, 10, 15 og 20 % i villakspopulasjonen henholdsvis 1.08, 4.52, 11.41 og 20.06 %. Selv om en moderat fraksjon av disse SNPene

[2] Referansegruppen påpeker at seleksjon virker på enkeltstående genvarianter, ikke på kombinasjoner av genvarianter, med mindre genvariantene er nært koblet. Uansett hva man måtte mene om dette, er det viktig å påpeke at det ikke har vært direkte seleksjon for redusert ferskvannsoppvandringsevne. Så selv om den fenotypiske endringen av oppdrettslaksen er kun basert på seleksjon på enkeltstående genvarianter, så er dette ikke i strid med den foreslåtte genetiske arkitekturen over. Vi sier kun at de rømtlaksene som fanges i elv har tilfeldigvis fått utdelt en genotypisk signatur som er betydelig mer lik villaks enn hva den største andelen av rømtpopulasjonen besitter, og de innehar derfor ikke en oligogen kombinasjon av alleler nødvendig for å forårsake en sterkt redusert ferskvannsoppvandringsevne.

er assosiert med dysfunksjonelle genvarianter av betydning for ferskvannsoppvandringsevne, er likevel antallet mer enn stort nok til at det er rimelig å anta at den polygene basisen er betydelig.
Selv om modellen er konsistent med resultatene betyr det ikke at den er riktig, og en sterkere validering av dens gyldighet vil kreve ytterligere tester som beskrevet under. Men formuleringen av en slik modell gir retning for et eventuelt videre arbeid, den åpner for konstruktiv diskusjon, og den gir grunnlag for en alternativ filtreringsprosedyre.

Den foreslåtte genetiske arkitekturen tilsier at en kan filtrere for SNPer fullstendig populasjonsuavhengig og derved relaksere kravet om at SNPene etter første filtreringssteg skal være felles for alle fire populasjonsgruppene. Dersom SNP-kandidatsettene som fremkommer er (delvis) forskjellige fra de SNP-settene som er brukt for å lage Figur 23, og at frekvensfordelingene i Figur 23 ikke endrer seg vesentlig, er dette en indikasjon på at forklaringsmodellen er riktig. Figur 24 viser samme type plot som Figur 23, men her er SNP-settene fremkommet ved at vi i filtreringstrinn 1 fravek kravet om at de identifiserte SNPene skulle være felles for alle populasjonsgruppene.

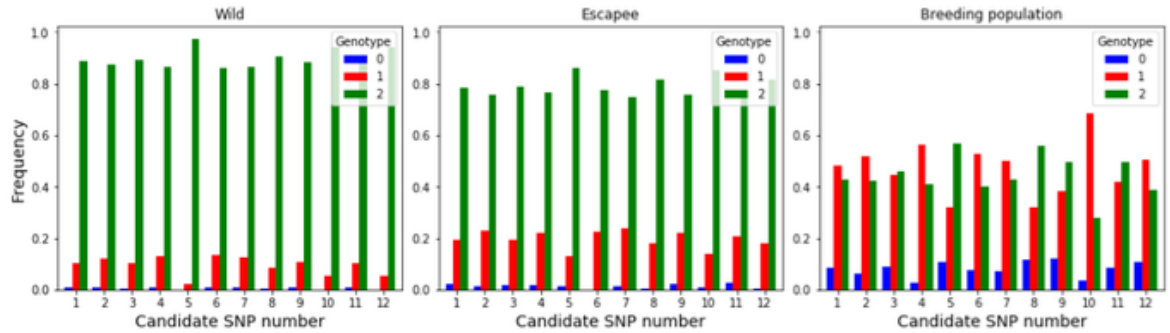


Figur 24: Frekvensfordelingen av Levenshtein-distansen mellom rømt oppdrettslaks og avlspopulasjon etter å ha justert for ulikheter som antas å ikke ha negativ betydning for ferskvannsoppvandringsevne i avlspopulasjonsindivider basert på en filtreringsprosedyre som ikke krevde at SNPene etter første filtreringssteg skulle være felles for alle fire populasjonsgruppene. Se tekst for ytterligere beskrivelse av hva som skiller denne figuren fra Figur 23.

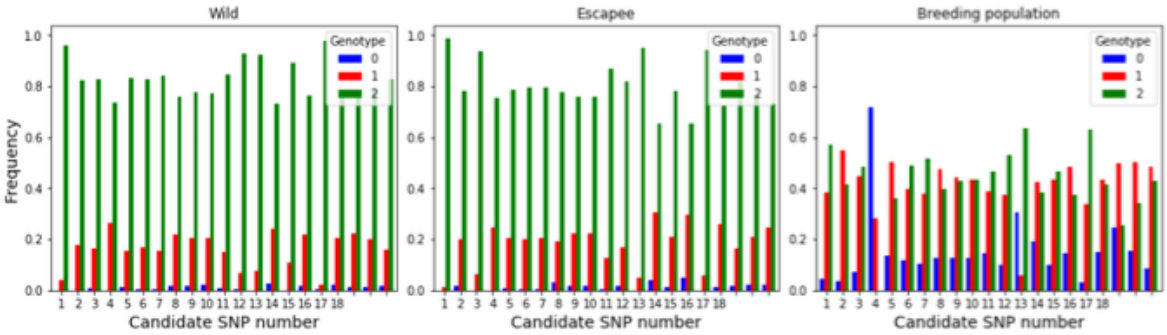
I filtreringstrinn 1 brukte vi EW = 7 og A1_max = 20 for alle populasjonene. I filtreringstrinn 2 kalibrerte vi AW-verdien slik at vi ble sittende igjen med om lag 100 SNPer for hver populasjonsgruppe. Ved å bruke AW_AquaGen = 23, AW_SalmoBreed = 20, AW_Mowi = 13, AW_Rauma = 21, ble størrelsene på SNP-settene henholdsvis 102, 97, 101 og 95. I dette tilfellet var en SNP felles mellom AquaGen og Rauma, en SNP felles mellom SalmoBreed og Rauma, 6 SNPer felles mellom SalmoBreed og Mowi, og 13 SNPer felles mellom Rauma og Mowi. Etter ytterligere filtrering basert på dekningsgrad-krav som beskrevet over satt vi til slutt igjen med 12, 21, 20 og 18 SNPer for henholdsvis AquaGen, SalmoBreed, Mowi og Rauma.

Vi ser at selv om Figur 24 skiller seg noe fra Figur 23, gir de nye SNP-settene, som forventet i henhold til prediksjonen, kvalitativt sett temmelig like distribusjoner. Det er grunn til å hevde at filtreringsprosedyren underliggende Figur 24 er den som bør brukes i et eventuelt oppfølgingsarbeid da kontrastene i

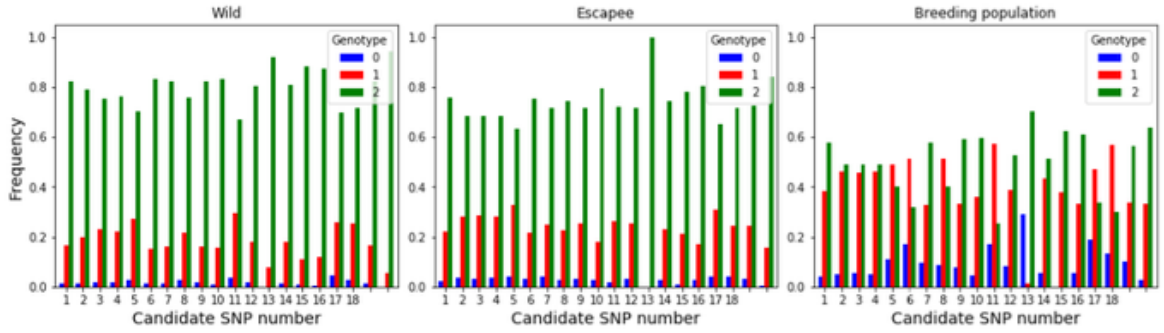
frekvensfordelingene til de ulike genotypene for hver SNP i kandidatsettene blir tydeligere enn om en anvender føringen at SNPene fra filtreringssteg 1 skal være felles for alle rømtpopulasjonene (Figur 25, 26, 27 og 28).



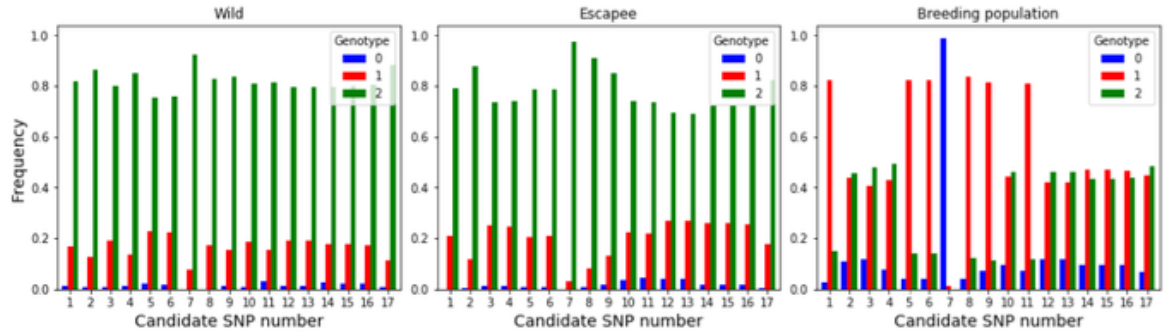
Figur 25: SNP-vis genotypefordeling i AquaGen hvor SNP-settet er basert på alternativ filtreringsprosedyre.



Figur 26: SNP-vis genotypefordeling i SalmoBreed hvor SNP-settet er basert på alternativ filtreringsprosedyre.



Figur 27: SNP-vis genotypefordeling i Mowi hvor SNP-settet er basert på alternativ filtreringsprosedyre.

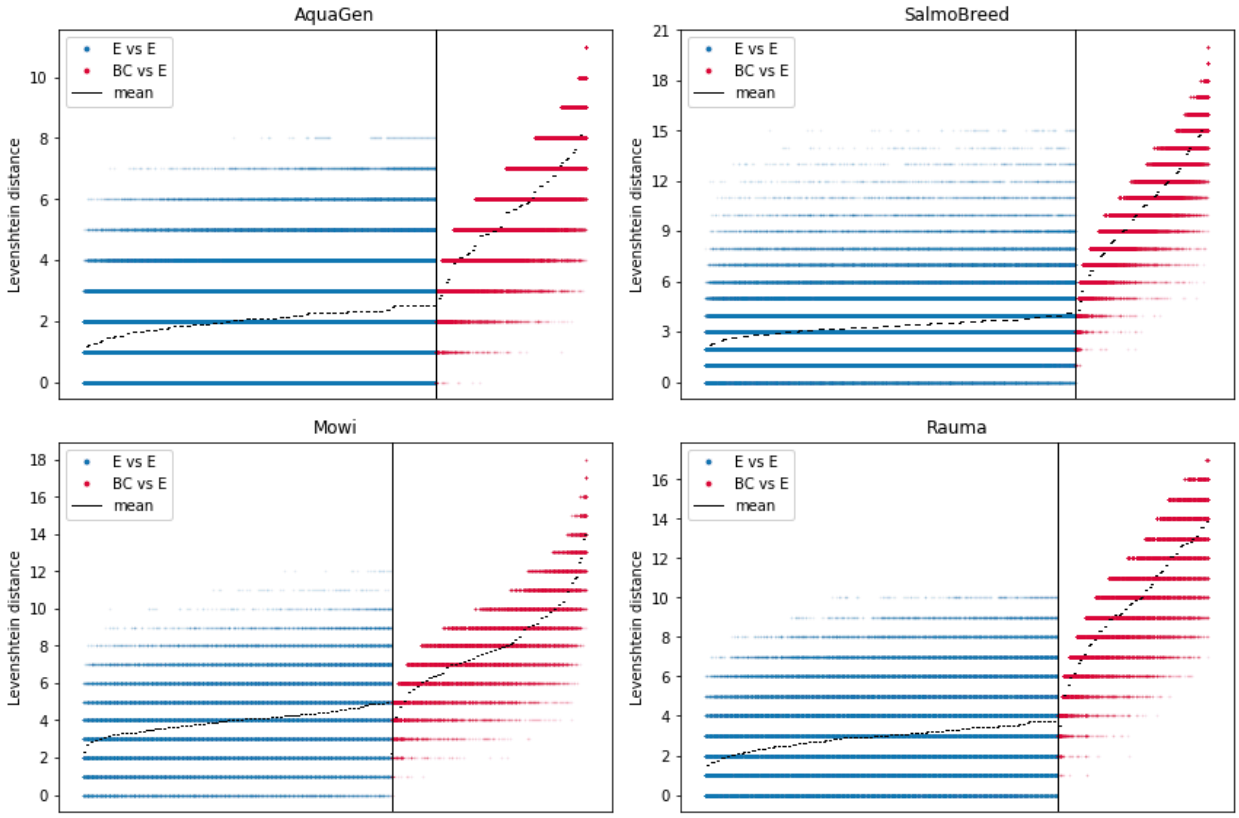


Figur 28: SNP-vis genotypefordeling i Rauma hvor SNP-settet er basert på alternativ filtreringsprosedyre.

Den foreslåtte genetiske arkitekturen forklarer hvorfor omfattende GWAS-studier med state-of-the-art programvare ikke ga oss resultater, da denne metodikken baserer seg på at enkelt-alleler har en synlig fenotypisk effekt i alle genetiske bakgrunner. I vårt tilfelle har vi en binær fenotype (oppvandret i elv eller ikke), det vil si at en ikke kan plukke opp små fenotypiske endringer som man kan gjøre med en kontinuerlig fordelt karakter. Dersom en dysfunksjonell genvariant må være sammen med andre dysfunksjonelle genvarianter i gitte kombinasjoner på andre loci for å påvirke karakteren, og det eksisterer mange slike oligogene kombinasjoner, vil GWAS-analysen nødvendigvis ikke greie å etablere assosiasjoner mellom genotype og fenotype.

Om forklaringsmodellen over er riktig så innebærer den at en har et betydelig antall frihetsgrader når en søker å sette sammen avlsdyrskrysninger som minimerer antall avkom med intakt ferskvannsoppvandringsevne. Dette kan gjøre at svekket ferskvannsoppvandringsevne lettere kan inkluderes som avlsmål.

Nødvendighets- og tilstrekkelighets-vurderinger: I Figur 29 har vi brukt den siste gruppen av SNP-sett og for hver populasjonsgruppe plottet gjennomsnittlig justert Levenshtein-distanse rømtlaksene seg imellom (sort linje i blått panel) og fordelingen av distanser underliggende hver gjennomsnittsverdi (blått), og gjennomsnittlig justert Levenshtein-distanse mellom avlspopulasjon og rømtlaks (sort linje i rødt panel) og fordelingen av distanser underliggende hver gjennomsnittsverdi (rødt). Gjennomsnittsverdien er basert på at en for en gitt sammenligning måler distansen mellom hvert individ i en gruppe med alle individer i den andre gruppen. En ser at det er påfallende lite overlapp mellom gjennomsnittsverdiene for de to sammenligningene i alle populasjonsgruppene.



Figur 29: Sammenligning av gjennomsnittlig justert Levenshtein-distanse mellom alle rømtlaks seg imellom (E vs E) og mellom opphavspopulasjon og rømtlaks (BC vs E), med tilhørende fordelinger underliggende hver gjennomsnittsverdi. Se tekst for ytterligere forklaring.

Den mulige praktiske verdien av Figur 29 er at den angir en metode som muligens vil kunne brukes som et verktøy i avl: **Om en sørger for at avkommet fra to avlsdyr (i en gitt avlspopulasjon for et gitt validert SNP-sett) har en gjennomsnittlig justert Levenshtein-distanse sammenlignet mot rømt oppdrettslaks fanget elv som kommer fra denne avlspopulasjonen, som er dobbelt så stor som den høyeste gjennomsnittverdien for den assosierte rømtpopulasjonen, så forventes det at avkommet vil ha redusert evne til å overleve i sjøfasen og/eller vandre opp i elv om det rømmer.[3]**

Med validert SNP-sett mener vi at resekvenseringsdataene har gitt indikasjoner på at SNPene er koblet til funksjonell genetisk variasjon med betydning for ferskvannsoppvandringsevne. Vi vil understreke at SNP-settene som er brukt til eksemplifisering ikke nødvendigvis er av en slik karakter, og at det selvfølgelig må gjøres mer eksperimentelt og analytisk arbeid før en kan avgjøre om en slik prosedyre er verdt å følge opp eller ikke (se under for utbrodering av dette).

5.0 Konklusjon

Den overordnede konklusjonen vi mener følger fra funnene over, er at de er verdt å følge opp om en ønsker å avklare om en genetisk tilnærming kan bidra til redusere problemene knyttet til oppvandring av rømt oppdrettslaks i ferskvann. En positiv avklaring betyr ikke at sluttresultatet blir at all ferskvanns-oppvandring vil opphøre om en omsetter kunnskapen i praksis, men de foreløpige resultatene indikerer at en vil kunne oppnå en vesentlig risikoreduksjon. Men vi vil understreke at resultatene på dette stadiet ikke gir grunnlag for å redusere nåværende innsats for å få ned rømmingstallene, fiske opp rømt oppdrettslaks, eller utnytte bioteknologiske metoder til å utvikle oppdrettslaks som ikke kan gyte i naturen.

6.0 Vurdering/drøfting av mulighetene for videre anvendelse av resultater fra prosjektet

Ut fra konklusjonen over foreslår vi et oppfølgingsprosjekt som vil fjerne de fleste usikkerhetene denne studien er beheftet med, og som vil bringe oss langt på vei mot en endelig avklaring om realiteten av funnene, og om de eventuelt kan la seg utnytte i praksis. De viktigste aktivitetene i dette prosjektet vil være:

- Genotyping av ytterligere 2000 rømte oppdrettslaks fanget i elv for å få større dekningsgrad av mulige genotyper som fremmer ferskvannsoppvandringsevne fra de enkelte avlspopulasjonene, og hvor en balanserer andelen tidligrømte (som har hatt havvandring) og seintrømte (som ikke har hatt det).
- Genotyping av 500 individer fra hver av avlspopulasjonene for å styrke representativiteten til datasettene.
- Bestemmelse, via genotyping og analyse, av den genetiske avstanden mellom avlspopulasjonene og produksjonspopulasjonene slik at en får et bedre mål for usikkerheten knyttet til bruk av avlspopulasjonene som oppdrettsreferanse. Genotyping av stamfisk vil kunne gi oss viktig informasjon, men referansegruppen påpeker at selv om en slik analyse kan gi en mer realistisk oppdrettsreferanse, vil variasjonen over rognbatcher kunne være så stor at en likevel risikerer å mistolke data. Dersom referansegruppens forslag om å sammenligne oppvandret rømt fisk mot ikke-rømt fisk fra de samme rognbatchene lar seg gjennomføre, vil dette sannsynligvis være en bedre strategi. Men en slik studie vil kreve betydelig assistanse fra avlsselskapene.

[3] Referansegruppen påpeker at «Denne seleksjonen må gjøres i sisteleddet, dvs. på foreldrene til salgrogna, i og med at ikke-additive effekter ikke er arvelige. Dette vil bli svært dyrt og svært begrensende for rognproduksjonen vår.» Vi deler bekymringen om at en slik seleksjon kan vise seg å bli uforholdsmessig fordyrende, men vi tror det er fornuftig å vente med å konkludere til en har et bedre kunnskapsgrunnlag vedrørende additivitet/ikke-additivitet, implementeringskostnader, og hvor begrensende en slik screening vil være for avlsdyrutvalget og rognproduksjonen.

- Øke sikkerheten til algoritmen som tilordner en rømt oppdrettslaks fanget i elv til avlspopulasjon (produksjonspopulasjon).
- Bruk av parallellprosessering i filtrerinsgsarbeidet for å kunne søke i et større genotyperom. Vi tror at tilgang til 30-50 regnekjerner over en rimelig kort tidsperiode vil være tilstrekkelig.
- For å øke sikkerheten vedrørende valg av SNP-er må en innhente sekvensinformasjon rundt kandidat-SNP-er for både villaks, rømt oppdrettslaks fanget i elv og avlspopulasjonene, slik at en kan bekrefte at de genene en predikterer å være kausale virkelig besitter genetisk variasjon som er konsistent med SNP-kandidaturet.
- Da det ikke kan utelukkes at genotypekombinasjoner som er hemmende for ferskvannsoppvandrings-evnen også kan ha negative effekter i oppdrettsmiljøet som sådan, ville det være ønskelig å analysere oppdrettslaks som dør i merd av uklare årsaker før de er salgsklare. Om dette er tilfelle vil resultatene kunne ha betydning for avlsarbeidet som sådan, men også være retningsgivende for hvordan redusere ferskvannsoppvandringsevnen gjennom genombasert presisjonsavl. En slik analyse vil kreve samarbeid med flere produksjonsselskaper.
- Vise at avlspopulasjonsspesifikke og godt validerte SNP-lister kan brukes av avlsselskapene til å bestemme avlsdyrkrysninger som vil minimere antall avkom som besitter genotyper som forårsaker opprettholdelse av ferskvannsoppvandringsevne, og at innarbeidingen av dette er økonomisk overkommelig og ikke vil negativt influere eksisterende avlsmål. For å kunne automatisere denne prosessen vil en ideelt sett trenge programvare som med betydelig grad av sikkerhet kan forutsi fordelingen av avkomsgenotyper basert på informasjon om genomstruktur (rekombinasjonsrater, koblingsinformasjon). Men i og med frihetsgradene i valg av SNP-er diskutert over, kan en sannsynligvis greie seg med betydelig enklere *in silico* modeller dersom en fokuserer på å sette sammen SNP-er som er i koblingslikevekt med hverandre.

Det er rimelig at avlselskapene, selv om en kan vise at implementeringen vil være gjennomførbar, vil kunne være avventende til å ta i bruk denne kunnskapen før en har uomtvistelig bevis for at den vil *de facto* fjerne eller i overveiende grad redusere problemet med rømt oppdrettslaks i elvene. Et slikt bevis vil kunne fremskaffes ved å gjøre et kontrollert utsettings- og gjenfangstforsøk, hvor en plukker ut to grupper av smolt fra hver av de fire avlspopulasjonene, den ene gruppen bestående av genotyper en forutsier vil vandre tilbake til elv, den andre gruppen bestående av genotyper en forutsier ikke vil komme tilbake til elv.[4] Det er mulig å gjøre slike forsøk i Norge – for eksempel på NINA Forskningsstasjon Ims med elven Imsa som er kontrollert med to-veis fiskefelle. Et telemetriforsøk på voksne oppdrettslaks med ulike genotyper vil også kunne anvendes for å fremskaffe et sterkere beslutningsgrunnlag.

Men før en kan gjøre slike forsøk må en gjennomføre det oppfølgingsprosjektet som er skissert over, slik at en kan velge SNP-kandidatsett med langt større konfidens enn vi kan gjøre i dag. Nå når både SNP-arrayet og analyseverktøyene er tilgjengelige mener vi at dette oppfølgings-prosjektet kan gjennomføres på noe over ett år om det gis tilstrekkelige ressurser.

[4] Referansegruppen mener at «det vil være svært vanskelig å lage slike grupper pga egenskaper virker veldig polygenisk og det er ingen praktiske måte å beregne avlsverdier med brukbar sikkerhet ut fra disse resultatene». De foreløpige resultatene visualisert i Figur 29 indikerer at vi vil være i stand til å sette opp slike grupper. Resultatene er ikke ment å legge grunnlag for å beregne avlsverdier. Dersom resultatene er i overensstemmelse med prediksjonene vil dette dokumentere at vi besitter operasjonaliserbare algoritmer som med betydelig sikkerhet kan forutsi om individer fra en og samme rognbatch har en sterkt redusert ferskvannsoppvandringsevne eller ikke. Dette vil også bekrefte at ferskvannsoppvandringsevne er en egenskap med en polygen basis, men med en oligogen realisering, som betyr at en ikke vil trenge å anvende genomisk seleksjon i utvalgsarbeidet.

7.0 Hovedfunn

- Det synes som det er klare genetiske forskjeller mellom oppdrettslaks fanget i norske lakseelver og de avlspopulasjonene disse rømte individene kommer fra.
- Oppdrettslaks fanget i elv fra fire ulike avlspopulasjoner har felles et betydelig antall SNP-loci hvor de genotypisk er mer lik villaks enn hva som tilsynelatende kan tilskrives tilfeldigheter.
- De genetiske forskjellene mellom rømt laks fanget i elv og avlspopulasjonene de kommer fra kan kobles til biologiske mekanismer som med stor sannsynlighet underligger laksens evne til å overleve i sjøfasen og vende tilbake til ferskvann etter rømming.
- Resultatene underbygger forklaringshypotesen som lå til grunn for initiering av prosjektet, nemlig at oppdrettslaksen har blitt selektert for egenskaper som indirekte har forårsaket at en stor andel av rømte oppdrettslaks har en sterkt redusert evne til å vende tilbake til ferskvann.
- Resultatene legitimerer en oppfølging som har som siktemål å produsere godt begrunnede lister av SNP-genotyper som kan danne grunnlag for kontrollerte utsettings- og gjenfangstforsøk for å bekrefte eller avkrefte om en ved genombasert presisjonsavl kan fjerne oppdrettslaksens evne til å vende tilbake til ferskvann uten å endre avlsmål for de fire avlspopulasjonene.

8.0 Leveranser

Detaljert oversikt over leveranser i prosjekt

- Gjennomført oppstartsmøte via Skype med prosjekt- og referansegruppen 6. mars 2017. Referat er tilsendt FHF.
- Gjennomført jevnlige møter via Skype for prosjektstyringsgruppen i perioden fra prosjektoppstart fram til sommeren 2018: 29. mars 2017, 9. mai 2017, 6. juni 2017, 22. juni 2017, 10. august 2017, 11. oktober 2017, 23. november 2017, 19. desember 2017, 16. februar 2017, 27 april 2018 og 18. juni 2018. Referater fra disse møtene er tilsendt FHF. Prosjektstyringsgruppen bestod av Kjetil Hindar, NINA; Sigbjørn Lien, NMBU; og Stig W. Omholt, NTNU. I tillegg var Kjell Maroni, FHF, informert om og invitert til alle disse møtene som avtalt.
- Ytterligere møter er blitt gjennomført for en utvidet gruppe kalt analysegruppen, da dette ble funnet mest hensiktsmessig for å sikre optimal fremdrift, fom november 2018 (da genotypingen var gjennomført) og frem til prosjektavslutning: 6. november 2018, 6 desember 2018, 15 januar 2019, 20. februar 2019, 4. mars 2019, 8. mars 2019, 14. mars 2019, 25. mars 2019, 2. april 2019, 12. april 2019, 30. april 2019, 8. mai 2019, 14. mai 2019, 27. mai 2019, 7. juni 2019, 14. juni 2019 og 21. juni 2019. Analysegruppen bestod av Kjetil Hindar, NINA; Sten Karlsson, NINA; Ingerid Julie Hagen Arnesen, NINA; Geir Bolstad, NINA; Sigbjørn Lien, NMBU; Nicola Barson, NMBU, Matthew Kent, NMBU; og Stig W. Omholt, NTNU.
- Statusrapportering til FHF per 31. august 2017 og 15. februar 2019
- Avviksrapportering til FHF av 15. februar 2019
- Månedlig rapportering til FHF per 22. mars, 1. mai og 1. juni 2019
- Faglig sluttrapport til FHF av 15. november 2019
- Administrativ sluttrapport til FHF av 15. november 2019

VEDLEGG:

Vedlegg 1: Oversikt over de 1980 analyserte individene av rømt oppdrettslaks (som Excel-fil).
Vedlegg 2: Liste over SNPer med høye diskordansverdier mellom 220 K array og 60 K array.
Vedlegg 3: Dokumentasjon av at genetiske forskjeller mellom rømt oppdrettslaks fanget i elv og opphavspopulasjonene deres kan knyttes til biologiske mekanismer assosiert med ferskvannsoppvandringsevne.