Breeding success and survival in relation to major effect loci affecting the age at maturation in Teno river Atlantic salmon (*Salmo salar*)

Jan Laine

MSc thesis
University of Turku
Department of Biology
4.5.2019

Degree Programme: MSc in Physiology and Genetics

40 ECTS

Referees:

1:

2:

Accepted:

Grade:

UNIVERSITY OF TURKU
Department of Biology
Faculty of Science and Engineering

Laine, Jan:

Breeding success and survival in relation to major effect loci affecting the age at maturation in Teno river Atlantic salmon (*Salmo salar*) *[Sukukypsyysikään merkittävästi vaikuttavien lokusten yhteys Tenojoen lohien (Salmo salar) lisääntymismenestykseen ja selviytymiseen]*

Master's thesis, 54 pp. 11 appxs.
Physiology and Genetics
May 2019

There is an age at maturity related trade-off between breeding success and survival in Atlantic salmon, and sex specific patterns exist in the expression of this life-history trait. A recent study identified three candidate genes, *vgll3, akap11* and *six6*, which had a major effect on this trait. Sex-dependent dominance was also observed in the gene *vgll3* with the strongest association to sea-age (years in ocean prior to maturation). This was assumed to be an adaptation to sexual conflict resulting from different optimal age at maturity between the sexes. However, direct effects on fitness were not investigated previously. Therefore, the aim of this study was to fill some of these knowledge gaps by studying sea-age candidate gene related breeding success and survival.

A dataset of 167 single nucleotide polymorphism (SNP) loci interspersed through the salmon genome including loci tightly linked with the sea-age candidate genes was generated for Teno river Atlantic salmon by sequencing DNA from one adult cohort and a subsequent cohort of juveniles on four consecutive years. Breeding success was studied by observing the trans-generational change in the sea-age candidate gene genotype frequencies and *vgll3* genotype related mate choice, and survival by observing the change of genotype frequencies in freshwater juveniles.

A significant deviation was observed between adult and juvenile *vgll3* and *six6* genotype frequencies. The homozygous *vgll3* genotype promoting later maturation in both sexes was significantly higher in juveniles than in adults, presumably due to better breeding success of males with this genotype. Homozygous *six6* genotype with opposing effect was similarly enriched. No *vgll3* related mate choice nor change in genotype frequencies among juveniles was observed. These results show that these two genes affect the breeding success and likely marine survival, but further studies are needed in order to assess the lifelong fitness effects of different genotypes.

KEYWORDS: Atlantic salmon, breeding success, life-history variation, SNP, sexual conflict, sexual selection, survival, Teno river

Contents

1. INTRODUCTION

Understanding the mechanisms of local adaptation, and the genetic architecture and evolution of locally adapted traits are among fundamental questions of evolutionary biology. The species commonly used to study these subjects may have been chosen either due to economic or conservation relevance or both (Ward 2000; Lerceteau, Plomion, and Andersson 2000; Meyer *et al.* 2010), or else because they are particularly suitable models for studying evolution (Merilä 2013; Mayden *et al.* 2007). All of these incentives come together in Atlantic salmon (*Salmo salar*), a culturally and economically important species that has a wide distribution on the northern hemisphere with high degree of genetic and phenotypic variation, but the wild populations are globally endangered or extinct in their natural habitats (Parrish *et al.* 2011). Along with overfishing and deterioration of natural habitats, climate change and the introgression of the gene pool by domesticated content |(i.e. by means of intogression of escaped farmed Atlantic salmon) may expedite the genetic diversity lost, and subsequently hamper local adaptation, hence survival in the wild (Parrish *et al.* 2011; Naish and Hard 2008; Jonsson and Jonsson 2009; Garcia De Leaniz *et al.* 2007). The growing demand for sustainable utilization of Atlantic salmon populations most of all have positioned the species alongside with other species of the Salmonidae family as a model species of evolutionary studies.

Accordingly, in this thesis, I investigated the fitness effect of a genomic region which have recently been identified to be a major genetic determinant of Atlantic salmon sea-age at maturity (Barson *et al.* 2015). The sea-age at maturity is a life history trait which has different optimal value (i.e. age) for females and males, causing a sexual conflict for the optimum trait value. Sex dependent dominance observed on this locus was speculated to be a resolution to this conflict, but it simultaneously raised questions about the degree of the resolution and whether this architecture is mutually beneficial for the sexes. Deeper understanding of this specific issue may help to assess and apply the best means for the conservation of local Atlantic salmon populations, enhance the farming and provide additional information about the possible evolutionary courses that the widespread phenomenon of sexual conflict may take among sexually reproducing species (Garcia De Leaniz *et al.* 2007).

1.1 Life history of Atlantic salmon S. salar

Atlantic salmon is a species from the Salmonidae family that inhabits the North Atlantic Ocean, freshwater lakes in North America and Europe and associated rivers of these water systems (Klemetsen *et al.* 2003). Atlantic salmon has a wide array of life-histories and huge variation in realization of these life-histories among different populations and individuals within a certain population (Fleming 1996; Klemetsen *et al.* 2003). Most Atlantic salmon populations are nevertheless anadromous: individuals spend 1-8 years their natal fresh water river after hatching, then after certain physiological and morphological changes, known as smoltification process, they migrate to the ocean, spends 1-5 years there feeding and growing and finally migrate back to their freshwater spawning areas to reproduce (Klemetsen *et al.* 2003; Saunders and Schom 2008; McCormick *et al.* 1998) (Figure 1). It has also been shown that in this migration process Atlantic salmon display high fidelity for their natal river, meaning that high percentage of the population tend to return to the same river to spawn that they hatched earlier in their life (Hansen, Jonsson, and Jonsson 1993). The combination of these different life stages (years spent in the river, years spent at sea, number of spawning attempts) make up the Atlantic salmon life-history.

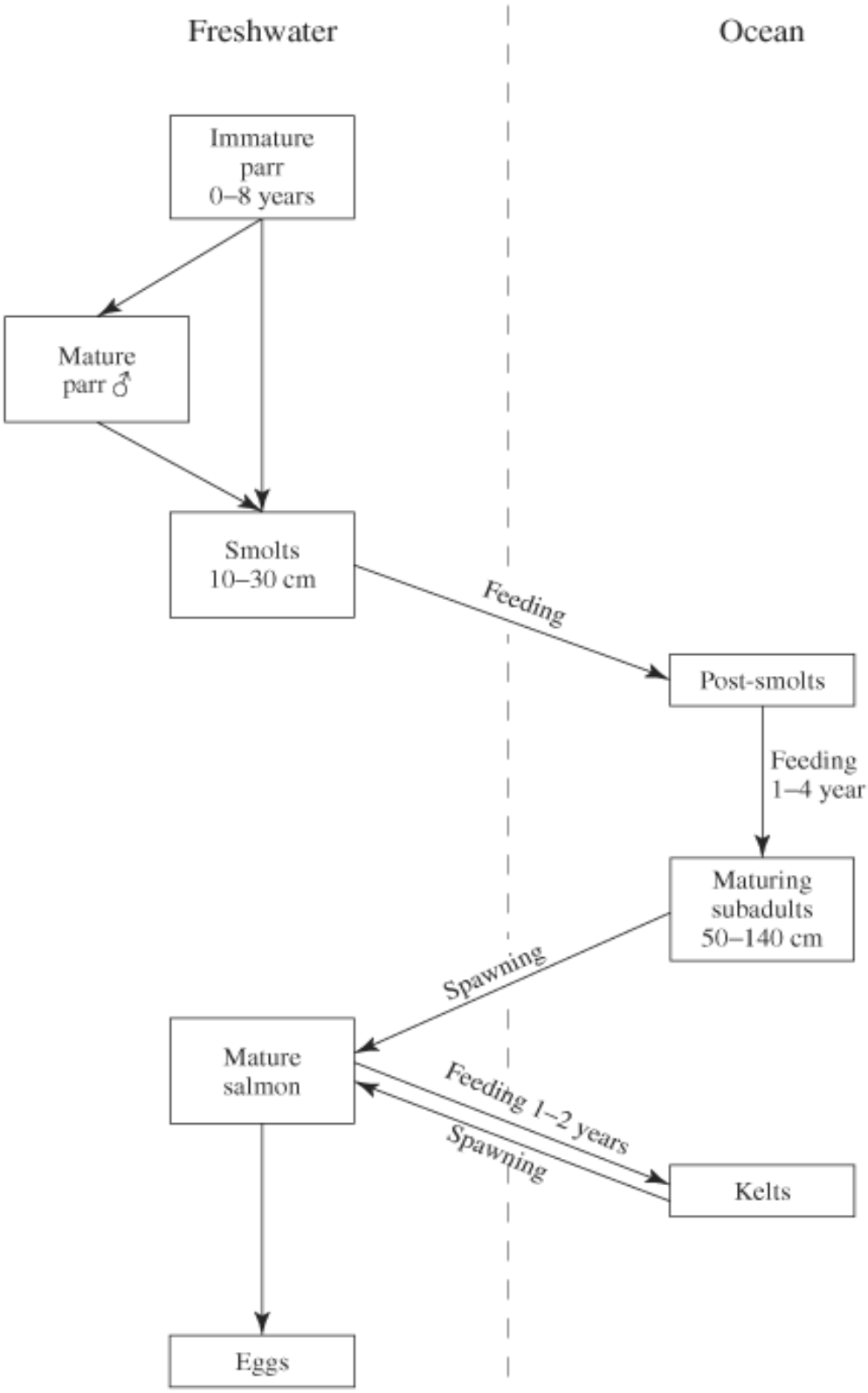

Figure 1. Major outline of life history of anadromous Atlantic salmon (Jonsson & Jonsson, 2009). Atlantic salmon juveniles mature after 1-8 years in their natal freshwater rive, migrate to Atlantic ocean where they feed and experience rapid growth until they mature after 1-5 years and return with high fidelity to their natal river for spawning. Some portion of male Atlantic salmon mature before smoltification and migration to sea as precocious parrs. Since Atlantic salmon is an iteroparous species, some individuals return to sea and participate spawning also later (these individuals are known as kelts).

The number of years an Atlantic salmon individual spends in the ocean, also termed as sea-age at maturation or simply sea-age, holds a pivotal role on an individual's reproductive success. As salmon have indeterminate growth, like most fishes, the longer the period spent in the ocean, the larger an individual will grow, and hence it has the potential for higher reproductive success through larger gamete number (also gamete size and quality in females) and territorial advantage compared to the smaller individuals during the spawning (Aas *et al.* 2011; Klemetsen *et al.* 2003; Heinimaa and Heinimaa 2004). On the other hand, a longer sea-period increases mortality risk before spawning and increases resource allocation for migration, gamete production and nesting or mate competition, which all compromises the opportunity to reproduce at all (Klemetsen *et al.*, 2003; Aas *et al.*, 2011). Thus, there is an evolutionary trade-off. This has led to emergence of distinct evolutionary strategies: some salmon mature early after one year at sea and return at a smaller size to the spawning grounds and thus reducing the early mortality risk, or alternatively they spend two or more years in the ocean, gathering more biomass and returning as more effective breeders, but in lesser numbers to reproduce.

1.2 Sexual dimorphism and sexual selection in Atlantic salmon

1.2.1 Underlying mechanisms driving sexual dimorphism and sexual selection

Since the multicellular organisms first developed, evolution has favoured gamete production of alternating degrees between two extreme ends of combinations of gamete size, motility and quantity (Parker 1978). On some occasions gamete production has fixated to these extreme ends resulting in production of few large stationary gametes, or alternatively larger quantities of smaller motile gametes, leading ultimately to the emergence of the two sexes (Parker 1978; Charlesworth and Charlesworth 1978). In many sexually reproductive species, at least some diversity beyond the gamete size is present between individuals of different sexes. These differences may be present in various trait types, such as in morphology, physiology, behaviour and life history, together which termed sexual dimorphism (Rice 1984).

Sexual dimorphism evolves by natural selection when the selection pressure differs between the two sexes (Lande 1980; Rice 1984). Sometimes this occurs due to ecological

causes i.e. when two sexes utilize resources somewhat differently (Shine 1989). However, empirical and theoretical evidences suggest sexual selection as the primary cause of sexual dimorphism (Lande 1980; Rice 1984). Sexual selection is a form of natural selection, where the selection pressure is imposed by other individuals of the same species and it acts on traits directly related to reproduction (Lande 1980; Kirkpatrick 1982). The pressure can be intersexual, where it occurs between the sexes and is based on e.g. mate preference, or intrasexual, where it takes place within one sex, and is mostly related to competition for mating opportunities and defensive behaviour over territories and mating partners (Lande 1980; Emlen and Oring 1977).

A fundamental aspect of sexual selection is that fitness of the offspring generation of sexually reproducing species is tied to the quality of the mating partners, and this is what ultimately gives rise to the sexual selection. One of the two sexes, commonly females, often invests more energy to gamete production and rearing of the offspring, and tend to be more selective towards its mating partners, which leads to assortative mating and to the intersexual form of the sexual selection (Lande 1980; Kirkpatrick 1982; Hunt *et al.* 2009). Correspondingly, the other sex subjected to this selectivity, commonly males, invests heavily to the intrasexual selection (that being principally competition over matings), which may be manifested, for example, as direct male-male competition, territorial defence, or development of ornaments and weaponry (Lande 1980; Hunt *et al.* 2009). Another important form of male-male competition and sexual selection is sperm competition, where the male gametes compete for female gamete fertilization following copulation, and at which point females can further act selectively by mediating the outcome of the sperm competition (Kekäläinen and Evans 2018). Besides mediating sperm competition, some other mechanisms may provide females the opportunity to influence which male gametes would eventually fertilize the ovum. This is termed cryptic female choice, and it may result in bias in paternity and the genotypes of subsequent offspring (Jennions and Petrie 2000).

1.2.2 Sexual dimorphism and sexual selection in Atlantic salmon

In Atlantic salmon, sexual selection acts heavily on life history traits (Tentelier *et al.* 2016; Fleming 1998, 1996). Females and males both have differing reproductive

strategies that they utilize, and both sexes display variation in their reproductive life history (Aas *et al.* 2011; Fleming 1998). One thing in common between sexes is that both males and females invest heavily to the spawning event; over 60 % of their total energy reserves (Aas *et al.* 2011; Fleming 1998). Especially migration has high energetic costs in anadromous life history strategies (Jonsson, Jonsson, and Hansen 1997). Besides the migration, in females, the energetic costs come from egg production, nest excavation, and nest defence (Fleming 1996; Aas *et al.* 2011). The number of eggs that a female can produce (i.e. fecundity) is tightly related to the female body size, and egg size similarly correlates with the body size (i.e. larger females produce more and larger eggs with higher protein content) (Fleming 1996; Aas *et al.* 2011; Heinimaa and Heinimaa 2004). Staying at sea longer, hence, is the most beneficial for females and ensures the highest fitness, since the more the eggs, the more offspring can be produced, and bigger eggs provides more resources for the offspring during the initial phase of their life history, thus potentially increasing the offspring survival (Thorpe, Miles, and Keay 1984; Aas *et al.* 2011; Fleming 1996). Larger size also allows females more efficiently excavate redds on better quality spawning grounds e.g. in faster flowing water with larger stones and better oxygenation, and more efficiently defend these nests from competing females (Aas *et al.* 2011; Fleming 1996). Given the rarity of smaller early maturing females in many populations (Barson *et al.* 2015), these benefits are apparently sufficient to counteract the negative effect on average fitness resulting from increased mortality risk due to multiple years spent at sea.

Similarly to the females, males also mainly express an anadromous life history. However, the sea period, on average, is shorter in males compared to females, likely because the body size isn't as imperative for male breeding success as in females (Fleming 1996, 1998; Barson *et al.* 2015). The more modest breeding success resulting from the shorter sea period and the consequent smaller body size is sufficient to allow utilization of this life history strategy and taking advantage of lower mortality risk at the sea, prior to reproduction (Fleming 1998, 1996). The overall trend in the Atlantic salmon is that males tend to mature more often after spending one winter at sea (1SW), while 2SW and 3SW fish are more common among females (Fleming 1996). Nevertheless, multiple sea winter males are not uncommon either, and these males tend to gain the dominant role on the spawning grounds and be simultaneously preferred by the females as a mating partners,

hence achieving the highest breeding success among the spawning males (Fleming 1996; Aas *et al.* 2011; Järvi 1990).

A second important reproductive phenotype expressed by Atlantic salmon males is precocious maturation as a parr, before the smoltification and the sea migration (Tentelier *et al.* 2016; Fleming 1996; Jones and Orton 1940). These males are substantially smaller than the anadromous males, and invest relatively more to gonad development and sperm production. Even though these precocious males can't establish dominance or compete in the ejaculate volumes with the anadromous males, they often manage to fertilize a certain portion of the eggs in the redds, by "sneaking" to the mating without taking the attention of the dominant male (Mjølnerød *et al.* 1998; Fleming 1996). The sperm of the precocious male parr is able to compete with the sperm of the anadromous males, and produce viable offspring (Kazakov 1981; Vladic and Järvi 2001). Hence, despite these individuals only have a minor breeding success, this life history strategy allows precocious parr to reproduce before the sea migration, substantially reducing the risk of mortality before the first reproduction event (Hutchings and Myers 1987; Fleming 1996; Aas *et al.* 2011). In other words, even if these males would later perish during the sea period of their life history, they have already had some contribution to the progeny of the next generation, which makes precocious maturation an evolutionary viable in males alongside the strictly anadromous life history.

1.3 Quantitative trait loci

Phenotypic traits of an organism are often so called complex polygenic traits i.e. they tend to express continuous variation and are product of environment, multiple genes, and the interaction within and between these factors (Lander and Schork 1996). Genes that are contributing to these polygenic traits with some particular effect size are called quantitative trait loci, or shortly QTL (MacKay, Stone, and Ayroles 2009). Effect of the QTL means the average change on the phenotypic trait that certain allele confers to the phenotype (MacKay, Stone, and Ayroles 2009). Depending on the trait and the species, the underlying genetic architecture of quantitative trait may involve a few large effect loci, or multiple more minor effect loci (Howe *et al.* 2003; Albert *et al.* 2008), whereby having many small effect loci and occasional presence of more major effect loci appear

to be a more common pattern in trait variance (Mackay 2001). The presence, location and effect size of the QTLs in organism's genome are studied via set of known genetic markers, and the phenotypic variation associated to these markers (MacKay, Stone, and Ayroles 2009; Barrett and Hoekstra 2011). These days, genetic markers most commonly used in these studies are single nucleotide polymorphisms (SNPs) (Coates *et al.* 2009).

1.4 Sexual conflict

Traits that express sexual dimorphism (e.g. body size, ornamentation and nursing behaviour), are mostly complex traits with a polygenic background (Lande 1980). This kind of genetic architecture may set constraints for sex specific evolution of sexually dimorphic traits, since most of the genes of a species are shared between the two sexes. Thus, a positive correlation between phenotype and genotype evokes a response in both sexes under natural selection (Lande 1980). However, the selection pressure may differ between the sexes and even be in different directions, which can lead to an antagonistic effects, where different extents of the variation of a trait are favoured differently between the sexes, and thus to balancing selection that restricts further sex specific evolution (Connallon and Clark 2014). In other words, if there is no refined resolution for the antagonistic effect, sexual conflict leads to a situation where neither sex can reach its optimal phenotype. Instead fitness of the population settles to a mean optimum, pulling both sexes away from their ideal optimal fitness.

Two forms of sexual conflict occur in the nature. First, the conflict may be interlocus, where effects of a single locus or several loci increasing unilaterally the fitness of one sex are suppressed by another locus or loci with similar effect in the other sex (Rice and Holland 1997; Chapman *et al.* 2003). The second form of sexual conflict is intralocus (Bonduriansky and Chenoweth 2009), and it occurs when different alleles of the same locus are differentially beneficial between sexes. When the subsequent phenotype resulting from the antagonistic locus is positively selected while expressed in the other sex, the relative fitness of the other sex lessens (Lande 1980; Bonduriansky and Chenoweth 2009). Thus the shared genome between the sexes constrains the sex specific evolution and neither of the sexes is able to reach its optimal fitness while the conflict prevails.

The sexual conflict may be partly resolved with sex specific gene expression or by genetic rearrangements that position the antagonistic loci to sex specific regions, such as the sex chromosomes (Ellegren and Parsch 2007; Rice 1984). With the presence of sex chromosomes, the inherited expression of the antagonistic alleles can be restricted partly or completely to the benefiting sex, and the other sex is relieved from the maleficent effects of these alleles (Rice 1984). Especially the X-chromosome is predicted to accumulate genes that code antagonistic fitness variation, and empirical evidence supports these predictions (Rice 1984; Gibson, Chippindale, and Rice 2002). However, other theoretical models have shown that in certain conditions autosomes may also maintain sexually antagonistic genes (Fry 2010). Such instances are possible, for example, when there is sex dependent variance in the dominance hierarchy of the antagonistic alleles (Fry 2010).

1.5 Genetic architecture of the age at maturity in Atlantic salmon

It was long known that sea-age variation is heritable and has a genetic basis in Atlantic salmon, and several recent studies aimed to identify genomic regions associated with the trait variation (Gutierrez *et al.* 2014; Johnston *et al.* 2014). In 2015, Barson *et al.* made a major discovery regarding the genetic architecture of the sea-age at maturity life history trait in Atlantic salmon. Using 208,704 SNPs genome-wide, and 57 European population they identified a major QTL affecting the trait on chromosome 25, explaining up to 39.4 % of the trait variation. The strongest signal in the region was in the vicinity of the gene *vgll3* (vestigial like family member 3), while another gene, *akap11* (A-kinase anchor protein 11), was located in the peripheries of the peak region. A SNP ~5kb downstream of the *vgll3* gene ($vgll3_{TOP}$ SNP) exhibited the strongest association with age at maturity, the alternative alleles of which differed in maturation age by, on average, 0.87 and 0.86 years in females and males, respectively (Figure 2). According to their phenotypic effect, alleles in this SNP were named *L* and *E* for late and early maturation, respectively. Furthermore, the alleles of this *vgll3* gene were shown to express sex specific dominance (Figure 2). Heterozygous $vgll3_{TOP}$ SNP females were more likely to mature later, similar to females with *LL* genotype, and male individual were more likely to mature earlier, similar to males with *EE* genotype i.e. the *L* allele was partially dominant in females, but the *E* allele was completely dominant in males. Since this facilitates the expression of the

more optimal phenotype in each sex, this expression pattern was speculated to be an adaptation to intralocus sexual conflict between the sexes (Barson *et al.* 2015; Fry 2010).

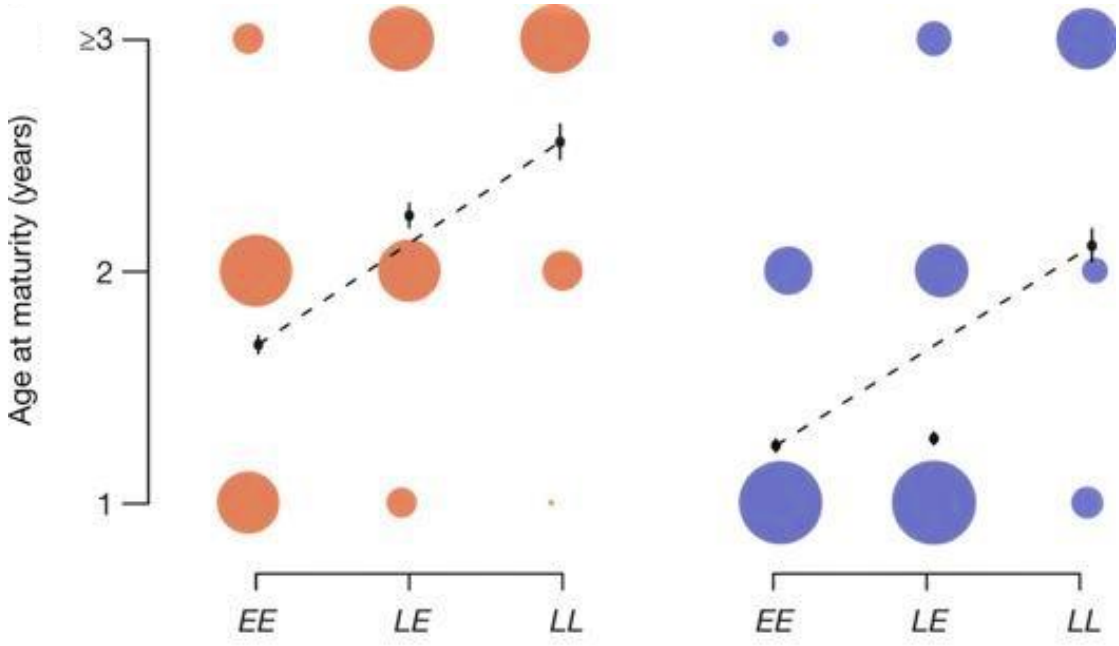

Figure 2. The effect of the *vgll3*$_{TOP}$ genotype to the age at maturity phenotype in the Atlantic salmon in females (red) and males (blue). The relative size of the circles represent the amount of individuals expressing sea-age phenotype in relation to certain *vgll3* genotype and the black dots predicted sea-age (Figure adapted from Barson *et al.* 2015).

A second region was found to be associated to the sea-age phenotype in the study of Barson *et al.* (2015), but only at the population level. This region was located on chromosome 9 and, among others, harboured a gene known as *six6* (SIX homeobox 6) (Jean, Bernier, and Gruss 1999). However, the association of this gene with the age at maturity trait wasn't statistically significant anymore after population structure corrections were incorporated in the analysis. Indeed, the *six6* genomic region was speculated to be locally selected, and further evidence has emerged to support this idea, though nothing conclusive have been shown and the interactions with the *vgll3* gene affecting the age at maturity remain similarly elusive (Pritchard *et al.* 2018). The *vgll3* gene is a transcription cofactor and has been shown to be negatively correlated with adiposity in mice (*Mus musculus,* Halperin *et al.*, 2013), a phenotype that is suggested to be an activator of sexual maturity in Atlantic salmon and in other fish species (Taranger *et al.*, 2010; Trombley, Mustafa and Schmitz, 2014). The *vgll3* gene is also linked to the

age at puberty in humans, indicating a conserved function in this locus (Cousminer *et al.* 2013). The *akap11* gene has a function in spermatogenesis, but it doesn't have any known conserved function to age at maturation or related traits (Reinton *et al.* 2000).

1.6 Objectives

Although Barson *et al.* (2015) inferred the sex-dependent dominance architecture to function as a mechanism providing resolution to sexual conflict, the actual effects of the different *vgll3*$_{TOP}$ genotypes and the linked sex dependent dominance on fitness were left unexplored, as was the breeding success of adults with certain genotypes. In this study, my objective was to fill some of these knowledge gaps by studying both the breeding success through sexual selection and survival in relation to different sea-ages and *vgll3*$_{TOP}$ genotypes. Similar effects of different genotypes of the two other regions associated to sea-age are also inspected. From now on, for the sake of convenience, the *vgll3*$_{TOP}$ SNP is referred just as *vgll3.* Likewise, the SNPs with strongest association to the sea-age linked to the *six6* and the *akap11* are referred with the names of these genes. The hypotheses of this study are listed in the end of this subchapter. The first aim of this study, to assess if there were differences in the breeding success resulting from sexual selection targeting the sea-age genotypes, is approached from two different angles. The second aim of the study is to assess the survival of the Atlantic salmon in relation to the sea-age linked loci identified by Barson *et al.* (2015), mainly focusing to the pre-smoltification phase in freshwater and to *vgll3.*

In order to satisfactorily evaluate the effects of certain sea-age genotypes on breeding success and survival, the studied population needed to be large enough to allow construction of robust dataset and such that the 1SW and the MSW phenotypes were present in both sexes. In this regard, the Teno river (Figure 3) was highly optimal: the river supports multiple Atlantic salmon populations in its tributaries that are throughout year habituated by juveniles and to witch adult Atlantic salmon of all sea-ages in both sexes ascend to spawn in relatively large numbers, especially in the main stream and closely associated populations (Johansen *et al.* 2016). Besides just the variety of phenotypes and the large population size, another advantage in the light of the aims of this study is the biased sex ratio of the spawning adults of the Teno river (Ellmén 2015;

Mobley *et al.* 2019). The operational sex ratio of Atlantic salmon is naturally male biased even in the populations with excess off females as the males can participate spawnings repeatedly whereas the females can only lay restricted number of eggs (Fleming 1998; Aas *et al.* 2011). Hence, the highly male biased sex ratio fortifies the effects of the biased mating system rendering the male-male competition extremely intense and is thus likely to bring forth all the possible effects of sexual selection. The exact separate aims of this study are listed below.

1.1) Trans-generational change in allele frequencies: I studied how the sea-age genotype frequencies transit through generations by comparing genotype frequencies between the spawning adults captured in one spawning area of the Teno River and their offspring several months after hatching. Significant differences in the genotype frequencies between the generations may provide an indication that some genotypes convey higher mating success or fitness than others during the spawning period. The null hypothesis was that the observed offspring genotype frequencies do not deviate from those expected under random mating conditions.

1.2) Assortative mate choice: The known mating pairs, based on the previously established pedigree (Ellmén 2015), were explored in order to detect possible patterns in the mate choice and assess if individuals of certain *vgll3* genotype and sex combinations prefer such mating partners with whom the produced progeny would have the optimal sea-age genotype. Since the sex dependent dominance is suggested to resolve sexual conflict, the heterozygous *vgll3* genotype is likely the most optimal due to heterozygous advantage that allows the different sexes to express the more advantageous phenotype. The null hypothesis was that mating does not deviate from random pairing. Besides observing just the mate choice, the possibility for segregation distortion resulting in offspring genotype frequencies deviating the Mendelian ratios was also explored in this context.

2) Juvenile survival in the natal fresh water river: The second aim of this study, assessment of sea-age genotype related survival in juveniles, was studied as change in the genotype frequencies between annually sampled Atlantic salmon juveniles from the same Teno River cohort in their natal freshwater river. Assuming there were no differences in straying rates between genotypes, changes

in these frequencies would reflect selection targeting certain genotypes during the fresh water life history phase. Since these genes are candidate genes for regulating age at maturation, a major event in individual's life history, they were assumed to govern or strongly contribute to crucial physiological mechanisms in Atlantic salmon (Fleming 1996; Aas *et al.* 2011). Hence, the physiological differences mediated by the alleles of these genes could already affect the overall physiology, growth rate and behaviour on earlier life history stage, and thus subsequently survival. Changes in the sea-age genotype frequencies were studied both between the different yearly sample groups and between the sexes of sample groups of the same year. The null hypothesis was that the sea-age genotype frequencies between these groups do not differ in the freshwater. Lastly, the comparisons of adult and juvenile sea-age genotype frequencies can allow some speculation of selection targeting the sea-age genotypes during the growth period at sea.

2. MATERIALS AND METHODS

The study material originates mostly from samples used in (Ellmén 2015) as well as juvenile samples collected in a similar manner in subsequent years. In fall 2011, anadromous adult Atlantic salmon were captured and sampled on a spawning region of the Teno River system on the lower Utsjoki tributary, close where it unites with the Teno main stream (Figure 3). These individuals were captured during a two week period preceding the initiation of the spawning at that site. The capturing was conducted with nets, from which the fish were immediately detached and moved to nearby pen. In addition to other phenotypic measurements taken from captured salmon, a fin sample was cut from the adipose fin for later DNA extractions. After these procedures the fish were marked for avoiding recapture and released back to the river allowing them to take part to the spawning. In total 54 adult salmon were captured that year.

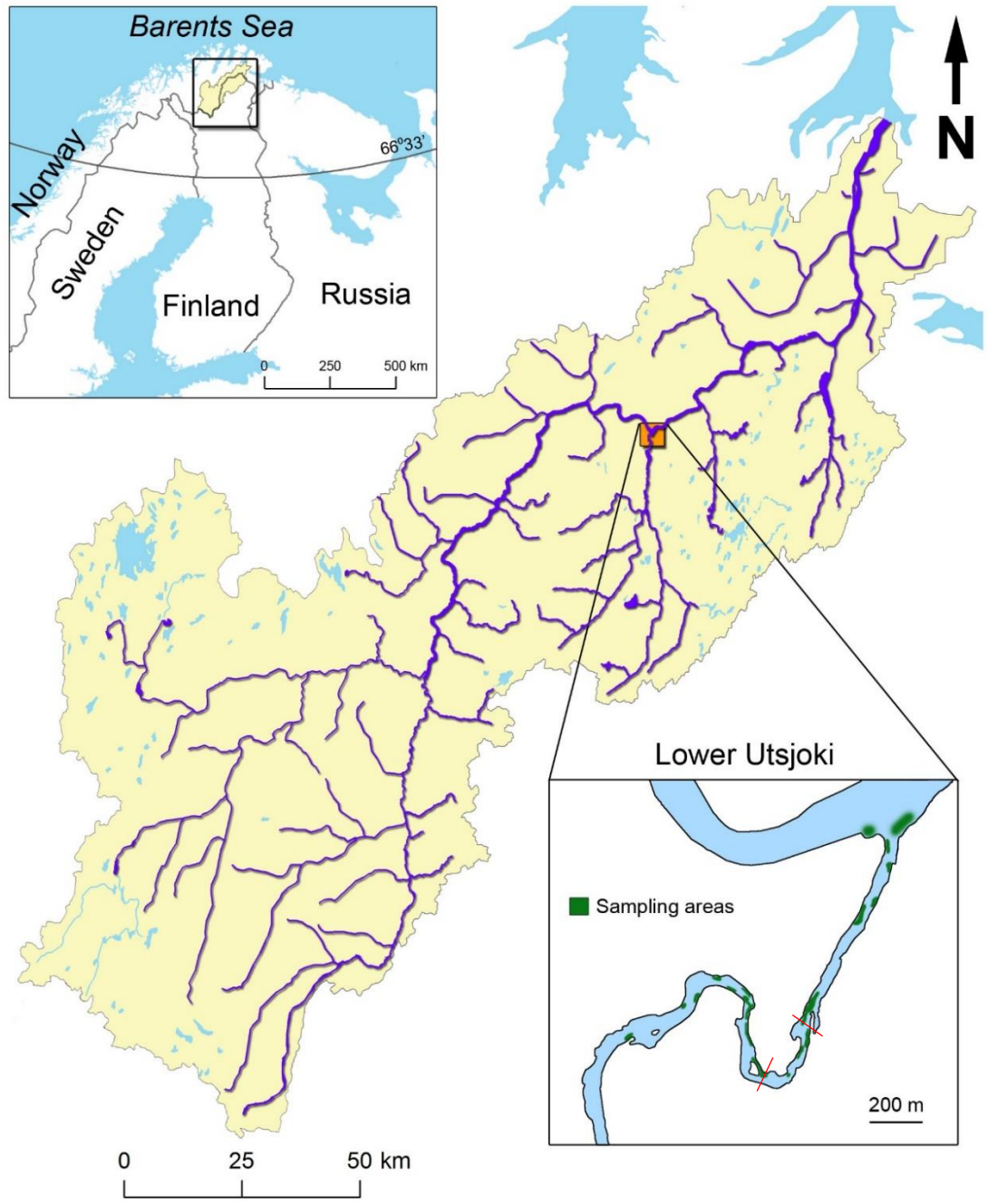

Figure 3. Teno river system bordering Finland and Norway and lower Utsjoki sampling sites. The whole Teno river system is coloured in purple, and the orange square marks the lower Utsjoki sampling site where Utsjoki meets and merges to the Teno mainstream. More detailed zoomed-in picture of the lower Utsjoki depicts in green the electrofishing sites of the juvenile Atlantic salmon, and the red lines marks the area of between which most of the future spawning adults were captured (49 out of the 54). Picture courtesy of Kenyon B. Mobley, modified from Mobley *et al.*, 2019.

In the following fall in 2012, the same region of lower Utsjoki was re-visited, this time to sample juveniles hatched that year (referred as +0y 2012) that were potentially fertilized by the adults sampled the previous fall. These juveniles were captured by electrofishing,

and the adipose, or in some occasions the anal fin, was cut for DNA extraction. In total 826 individuals were captured, and together with the adults from 2011 these samples constituted the sample material in the Ellmén 2015 study. However, similar juvenile captures and samplings were continued also on the following falls of 2013, 2014 and 2015 as an ongoing material collection practice for related future and more in depth studies, giving rise for additional +1y 2013, +2y 2014 and +3y 2015 sample groups of different ages from the Atlantic salmon cohort that were fertilised by the 2011 adults and had hatched in 2012. In total 911 1+, 184 2+ and 75 3+ juveniles were captured on the years 2013, 2014 and 2015, respectively.

For his study, Ellmén extracted DNA from the 2011 adults and +0y 2012 juveniles and genotyped these individuals using 14 microsatellite loci. From this microsatellite data, he constructed a pedigree in order to study the local adaptation advantage in reproduction, where the parentages of the spawning adults from the 2011 to the +0y 2012 juveniles were assinged. Based on this pedigree, 14 mating pairs could be identified among the spawning adults of 2011, and all together 54 offspring resulted from these pairings. In my study, I utilized the sample material of 2011 adults and +0y 2012 collected by Ellmén together with later collected +1y 2013, +2y 2014 and +3y 2015 samples. DNA from all but 32 of the +3y 2015 samples was already available. I extracted the DNA of the remaining 32 samples. I analysedall these samples with a SNP based genotyping assay developed for Atlantic salmon, that includes among others the *vgll3*$_{top}$ SNP and other sea-age candidate gene marker SNPs (Aykanat *et al.* 2016). Using the gained genotype information of these samples together with the pedigree established by Ellmén, I was able to address my previosly stated hypotheses.

2.1 DNA extraction

I extracted the DNA from the remaining 32 +3y 2015 juvenile samples with a QIAamp DNA Mini Kit (www.qiagen.com) following the manafacturers instructions. An equal sized piece of (1-2 mm) was cut from each of the collected adipose fin samples for the DNA extraction, and the extracted and purified DNA was eluated to110 µl of MΩ.cm Milli-Q® water.

2.2 Sequencing workflow

The sequencing pipeline used was introduced by Aykanat *et al.* (2016). The protocol had since been optimized so that the number of SNP targeted by sequencing was reduced from original 211 to 197. In total 196 primer pairs were utilized, with each pair amplifying a genomic DNA region spanning a targeted SNP site. Further minor optimizations were done during the study to factors such as primer concentrations, number of the multiplexes used, PCR-protocols and methods for amplicon purification after the PCR-2. The final conditions used during my thesis are presented in Appendix A, as they represent the most optimised conditions.

2.2.1 Ion Torren PGM platform and DNA libraries

The sequencing of the amplified SNP sites was done on the Ion Torrent PGM platform, a next generation sequencing platform based on a sequencing by synthesis approach, where incorporation of nucleotides to elongating DNA strand are detected in real time *via* a pH change induced by hydrogen ions released in the synthesis reaction (Rothberg *et al.* 2011; Goodwin, McPherson, and McCombie 2016). The actual sequencing reaction on this platform takes place on a dedicated Ion Torrent chip, build around a CMOS sensor that is widely used in the present day's technology and thus relatively cheaply available, coupled with ion-sensitive field-effect transistors (ISFET) that allow the detection of the released hydrogen ions. This platform was chosen, since it's adequate for rapid sequencing of few hundred base pairs long DNA strands simultaneously from multiple loci and individuals, such machine was available in house and protocols for desired sequencing workflow already existed (Aykanat *et al.* 2016).

The design of the Ion Torrent chips set the framework for the workflow for the preparation of the DNA libraries that were sequenced, since the chips can only produce a certain number of reads, and a balance between the number of individuals included and the sequenced loci was needed in order to the most cost-efficiently ensure enough reads for accurate genotyping. In total nine Ion Torrent compatible DNA libraries were prepared and sequenced. As the amount of SNPs sequenced stayed constant, the amount of individuals included in each library varied from 40 to 384, and the Ion Torrent sequencing chip scale was selected accordingly. Most libraries included samples from three 96-well

plates (around 300 samples) and were sequenced on Ion Torrent 318 v2 chip, which yielded adequate coverage for each locus of each individual from libraries up to 384 individuals, but also Ion torrent chips 314 v2 and 316 v2 were used for the smaller libraries for optimal cost to performance ratio.

The preparation of the DNA libraries was performed in several steps (Figure 4). As pre-extracted DNA was available from majority of the samples, the initial step of the library preparation was amplification of the targeted SNP containing loci. Later in other PCR the amplified marker loci DNA was further amplified with primers containing sample specific barcodes, thus allowing pooling of all sample DNA together, while still retaining the means to differentiate the sample DNA based on its origins. These two PCRs were termed as PCR-1 and PCR-2, respectively. Preceding the sequencing, a further PCR was implemented, so called emulsion-PCR, for the DNA library under construction. This third PCR transformed the library DNA in a form compatible with the Ion Torrent platform *via* adapter sequence containing primers and other reaction components permitting the sequencing reaction to take place later on an Ion Torrent chip. In addition to the PCRs described here, adjacent steps were included to the protocol, such as purification of the PCR products, concentration measurements and pooling of the sample DNA. An overview of the general workflow is presented in the figure 4.

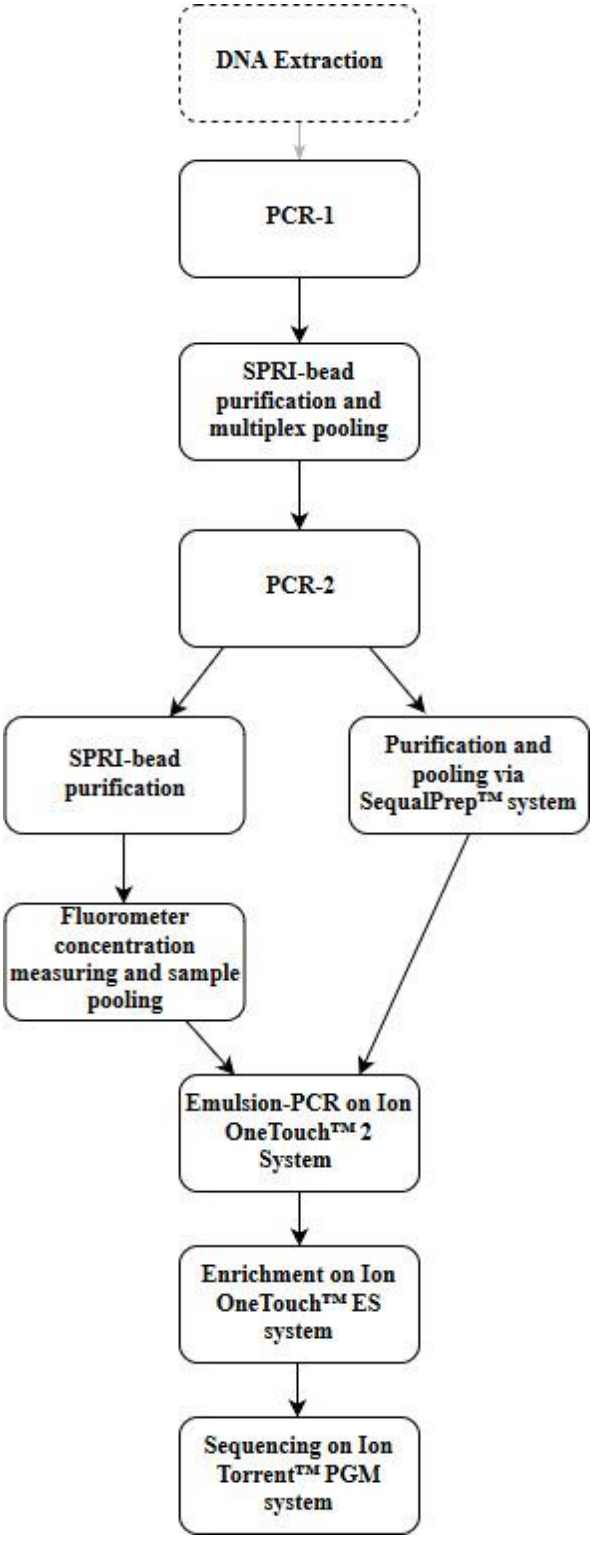

Figure 4. Workflow for DNA library preparation and sequencing. At the branching point alternative protocols for the workflow are presented. DNA extraction was necessary only in case of few samples and thus it wasn't part of the standard workflow.

2.2.2 Primer pool preparation and PCR-1

In preparations for the PCR-1 all SNP primers were pooled together so that the final primer concentration in the reactions were either 0.2, 0.1, 0.05 or 0.025 µM, depending on the amplification efficiency of the primers in question. PCR-1 was also divided to different multiplex modules where different set of primers were separated to different PCR reactions. These primer concentrations and groupings were made in order to equalize and optimize the amplification of each key SNP loci so that the final sequencing reaction would produce sufficient coverage for each locus enabling precise genotyping. This also allowed slight enhancement of amplification of the crucial sea-age marker loci.

For most libraries, a preparation workflow including two multiplexes at the PCR-1 stage was implemented. In some instances, protocols with four and one multiplexes were also utilized. The four multiplex protocol was used on the early optimization stages, but abandoned in favour of two multiplex protocol that was found to work equally well and more efficiently. In contrast, one multiplex protocol was not pursued in which some unspecific primer pairs produced sizeable portion of unspecific product reducing the sequencing efficiency drastically (6^{th} library prepared). In order to avoid the risk of further compromising the quality of the data produced, the two multiplex protocol was adopted on the later library preparations.

2.2.3 SPRI-bead purification

Following multiplex PCR-1, the amplicons from each multiplex of each individual were pooled together in equal volumes, thus bringing together all SNP markers. This pool was then subsequently treated with SPRI-bead purification method (solid phase reversible immobilization beads, Sera Mag, GE Healthcare Life Sciences; www.gelifesciences.com) in order to free the amplified markers from unincorporated primers, primer-dimers and other products formed during the PCR-1 due to non-specific amplification (Brownie *et al.* 1997). SPRI-bead solution was employed here in the ratio of 1.8 in reference to the pooled PCR-1 products, in order to maximally reduce the amount of non-specific amplicons present without compromising the marker loci fragments amplified. The purification process was carried out following the manufacturer's instructions.

2.2.5 PCR-2, purification and pooling

To enable the pooling of all the markers from each individual of the same DNA library, the purified PCR-1 product was amplified applying Ion Xpress™ Barcode forward primers (www.thermofisher.com) together with customized barcode containing reverse primers in the second PCR phase. The use of forward and reverse barcodes in combination allowed formation of a unique barcode for each individual, thus to individually match sequence output to an individual. The Ion Xpress™ forward primers consisted of 96 primers with unique barcode sequence, and the customised reverse barcodes consisted similarly of 8 primers with unique barcode sequencing. The barcode of the forward primers allowed to resolve the well of origin on each 96-plate for each sample, and the barcode of the reverse primers allowed the identification of the 96-plate that the sample was from. The forward primers also contained an Ion A adapter sequence, that were later used to bind the library DNA to the Ion Torrent sequencing system.

Two different PCR-2 protocols and immediate downstream methods were used in this study as a part of the on-going optimization process. The PCR-2 method applied on the most cases of library preparations included a second SPRI-bead purification, equivalent to the purification of the PCR-1 products, followed by quantification of the ds-DNA of each cleaned sample via Qubit 2.0 fluorometer (Invitrogen; www.invitrogen.com). The SPRI-bead solution to library ratio used at this point was 1.4 due to the added length of barcode and adapter sequences. Furthermore, since the PCR-2 reaction was conducted in one-tube without separate multiplexes, extra H_20 was added to ensure reasonable working volume for the SPRI-bead cleaning. Downstream steps in the second SPRI-bead protocol were otherwise the same as in the case of PCR-1. Following the cleaning of the targeted amplicons, a Qubit 2.0 fluorometer was utilized to measure the DNA concentration of each individual sample, and this concentration was further used to interpret the volume to use to achieve (40 ng) of DNA from each sample in the final pooled library.

The second PCR-2 protocol implemented in the library preparation was a faster, less labour demanding method that did not require SPRI-beads nor Qubit 2.0 fluorometer measurements. Instead, SequalPrep™ Normalization Plates and related protocols were applied (www.thermofisher.com). The wells of the 96-plates of this normalization system

change their charge depending on the ambient pH, allowing selective binding and releasing of negatively charged DNA. As a result, short non-specific amplicons that are less capable to bind can be discarded when longer targeted DNA fragments more readily bind to the surface of the well. For the proper execution of this protocol, enough DNA needs to be present at the initial binding step. To achieve this, larger reaction volume (20 µl) and extra amplification cycles were added to the PCR on this protocol. The complete purification process with the SequalPrep™ system was carried out following the manufacturer's instructions. The end product of this system was the pooled and purified DNA library without any additional steps required.

2.2.6 Library sequencing

After pooling all of the sample DNA together as a one DNA library, the library was made compatible with the Ion Torrent PGM sequencing platform. To achieve this, Ion OneTouch™ 2 System (www.thermofisher.com) was utilized to perform an emulsion-PCR, where library DNA is bind to and amplified on top of Ion Sphere Particles (ISP). These particles are bead structures that are covered with sd-DNA adapters compatible with the Ion A adapter -sequence incorporated to the library DNA sequences during the PCR-2. Thus, these particles were able to bind the library DNA and link it to the sequencing platform later when the ISPs were imbedded to the micro wells on the Ion Torrent chip. In the emulsion-PCR millions of reaction centres are created within the reaction oil so that only one strand of library DNA is coupled with one ISP in the reaction centre. When the library DNA was then amplified on the ISP, one specific particle was covered with clonal DNA of one strand of the library DNA. This was essential for creating coherent signal during the final sequencing reaction. Ion PGM™ Hi-Q™ View Sequencing Kit was used when the emulsion-PCR and the Ion Torrent PGM sequencing was carried out.

To create conditions where only one strand of DNA and one ISP are coupled in a reaction centre in the emulsion-PCR, the concentration of the library DNA needed to be highly optimized: extremely diluted but high enough to be sufficiently amplified (8 pM here). In order to achieve such precise concentration 2100 Bioanalyzer Instrument (www.genomics.agilent.com) was utilized. The microfluidic technology of this

instrument enabled more accurate determination of the concentration of the pooled post PCR-2 library compared to the fluorometer technology. Four to five replicated measurements were conducted from each library, extreme values were excluded and the average concentration was calculated from the remaining values.

Following the emulsion-PCR, the ISPs covered with clonal library DNA were enriched in the output from the Ion OneTouch™ 2. This enrichment was done on a One Touch ES station in a fully automated process (www.thermofisher.com). After the enrichment the actual sequencing could be executed on the Ion Torrent PGM machine. The machine was prepared and an Ion Torrent chip of appropriate capacity was loaded with the library DNA containing enriched ISPs by following the manufacturer's instructions.

2.3 Raw data processing and filtering

The raw signal data is heavily processed already in the Ion Torrent PGM parallel to the sequencing, and the final output data is provided as fastq-files. This file format holds both the raw nucleotide sequence of the read and the phred quality score measures for each nucleotide in the sequence. Since the Ion Xpress™ Barcode forward primers were used, the algorithms in the Ion Torrent PGM machine could already recognise these barcode sequences and divide the reads in different output files based on them. The further dissection of the reads in these files based on the custom reverse primer barcodes and read assignment to specific individual was done as part of the analysis pipeline build in the R programming language (version 3.3.2) (Aykanat *et al.* 2016). Similarly, the SNP identification in each read and final genotyping was done in the same R code pipeline (Appendix H). The variants of barcodes and SNP sites were identified by comparing the sequences of each read to a master table file working in tandem with the pipeline code and containing sequence information of all SNP site and barcodes involved (Appendix G & H).

The genotypes were called based on the ratios that each allele of a certain locus was amplified in certain individual (Figure 5). If the ratio of allele 1 compared to allele 2 was higher than 10.0, the individual was genotyped as homozygous for allele 1. Likewise, if the same ratio was lower than 0.1 the genotyped was called as homozygous for allele 2.

Finally, if the ratio settled between 0.2 and 5.0, the individual was genotyped as heterozygous. In case the ratios between the alleles fell outside these thresholds, the genotype was considered as inaccurate and was not called. In addition, if the phred score at the specific SNP nucleotide was less than 20, the genotype was not called. Also, if the overall coverage of the locus was less than 12, the genotype was left uncalled.

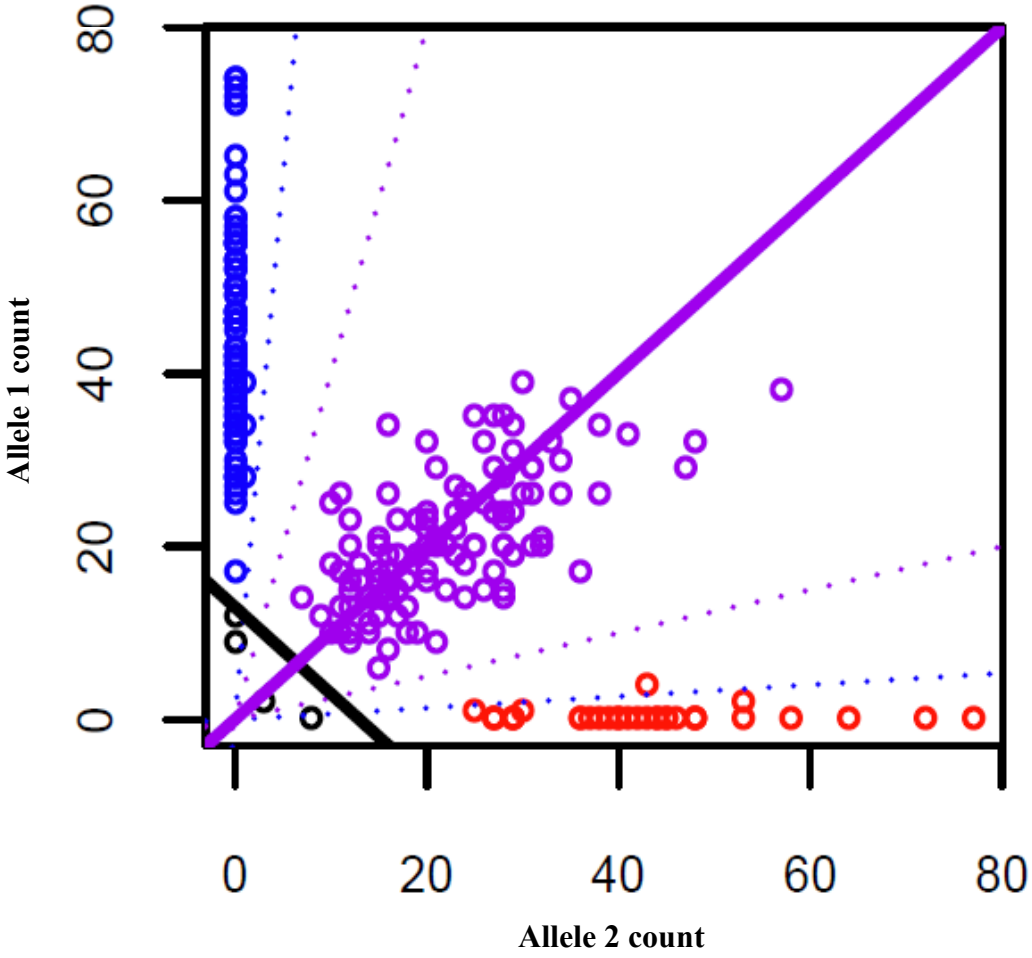

Figure 5. Example of genotype calling at a certain locus based on the coverage of each allele. Here presented are the *vgll3*$_{TOP}$ locus genotypes of the individuals included in the 9th Ion Torrent library. The blue circles represent the individuals genotyped as homozygous for the *L* allele, the purple circles represent the heterozygous individuals and the red circles individuals homozygous for the *E* allele. The black circles represent the individuals with uncalled genotypes due to low coverage. The dotted lines represent the thresholds used for genotype calling and the black traverse line represents the threshold under which the genotypes were not called.

The genotypic sexing was done based on the presence or absence of the sex marker locus. Since this locus is solely present in genomes of the male individuals, all individuals where this locus amplified with coverage close to the average coverage of other loci were assigned as males, and individuals where the coverage was zero or close to zero were assigned as females. However, certain fluctuation was allowed, as the amplification

success of the sex marker locus in males could be somewhat lower than the average coverage of other loci but still remain reasonable, whereas some level of coverage could be present in females due to a slight contamination. Thus, all individuals with sex marker locus coverage to average coverage of other loci resulting in ratio over 0.4 were assigned as males and individuals with ratio less than 0.15 were assigned as females.

As not all loci amplified adequately in all individuals, further filtering criteria were applied. Only loci that were genotyped in more than 80 % of the individuals were included in the final dataset. Similarly, only individuals whose loci were successfully genotyped over 70 % of the cases were kept in the final dataset. These measurements ensured that individuals and loci most susceptible for errors in genotyping would not introduce errors further downstream in the analysis.

2.4 Statistical testing

Changes in genotype frequencies between two Atlantic salmon generations and among classes within a juvenile cohort were studied here. Since genotype frequencies of two groups on a single locus form a conventional contingency table, statistical testing based on this method (utilization of contingency table) was applied to detect significant deviations in the genotype frequencies and hence address the hypotheses presented earlier. Contingency table was also constructed using expected and observed offspring genotype counts while studying the *vgll3* and s*ix6* related breeding success in the framework of trans-generational genotype frequencies. The expected offspring genotype frequencies were derived by calculating the likelihood of a certain allele from one sex to co-segregate with another allele from the opposing sex assuming completely random segregation, and this likelihood was multiplied with the total offspring count. Unconditioned two-tailed Fisher's exact test was used to test statistical significance. This test was preferred over e.g. chi-squared test due to its more conservative nature, as significance of differences of genotypes at multiple loci between multiple individuals is setting that is susceptible for false positives.

While comparing the trans-generational genotype frequencies, a further robust correction for genetic inflation was conducted by estimating and dividing the alpha thresholds for the P-values yielded by the Fisher's exact test with the genomic control factor λ (Devlin

and Roeder 2004). The estimated value for λ was calculated by dividing the median quantile of the chi-square distributed observed P-values (i.e. obtained from all SNPs used in the study) with a median quantile of a random chi-square distribution. Since the genotype, not alleles, were used here to construct the contingency tables, chi-square distribution with two degrees of freedom was used (Clarke *et al.* 2011). Genetic inflation could be present in the data due to sample duplications and other technical biases as well as population stratification (Yang *et al.* 2011). Q-Q (quantile-quantile) plots were constructed for a visual representation of the genetic inflation in the data as well as detecting significant deviations in the genotype frequencies that ranked in more extreme quantiles than expected in random distribution by plotting the observed P-value distribution against expected null distribution. The null distribution here was chi-square distribution with two degrees of freedom.

Fisher's exact test was also applied for testing assortative mate choice. In this case the contingency table was constructed so, that the genotype numbers of males and females formed the rows and columns in the table, respectively. The values in each cell of the table then represented the amount of known mating pairs of the two genotypes in question. The segregation distortion in the allele segregation was also tested in this context with Fisher's exact test by comparing the genotype frequencies of the offspring born for a known mating pairs to the expected Mendelian ratios.

2.5 Additional Utsjoki 2011 adult data

In the final dataset, 46 extra adult individuals were added to the dataset in order to achieve wider take of the gene pool of spawning adults of 2011. These adults were part of the same 2011 adult population of lower Utsjoki region as the 54 original adults with assigned parentages. Nevertheless, these specific 46 fish did not take part in the spawning, since they were not released post-capture. They were captured in the mainstream of Teno river and assigned to the same population as the fish sampled by Ellmén (2015) on the lower Utsjoki region, using molecular markers (Czorlich *et al.* 2018). These 46 adults were part of a similar SNP data of another study genotyped with the same platform and methods in the same laboratory, and were thus available for this study (Czorlich *et al.* 2018).

3. RESULTS

3.1 Genotyping success and sexing

3.1.1 Overall genotyping success

After filtering out 37 individuals with low geotyping succes (< 80 %), and 30 loci with low coverage (< 12x) or unreliable topography (inspected by eye. e.g. see Figure 5), 2059 individuals and 167 loci remained in the dataset for further analysis (Table 1).

Table 1. Genotyping success of the samples included in the study by sample groups. All individuals with 70 % of loci successfully amplified and genotyped were considered as successfully genotyped.

	+0y 2012*	**+1y 2013**	**+2y 2014**	**+3y 2015**	**Adults 2011**	**In total**
Initial data	826	911	184	75	100**	2096
Final data	803	900	182	74	100**	2059
Genotyping success	97.2 %	98.9 %	98.9 %	98.7 %	100 %	99.5 %

* The first part of the name of the sample group refers to the age as captured, and the second part to the year of capture.
** 100 individuals comprised of 54 that were captured by Ellmén (2015) and the 46 that were assigned to the same population with molecular markers and added to dataset to estimate allele frequency of population with less sampling error (Czorlich *et al.*, 2018).

Of the 2059 individuals 2038 (99.0 %) were successfully assigned to either sex by the genotypic sexing marker. All adult fish were successfully sexed, and genetic sexing was in concordance with the previously established pedigree, and phenotypic sex. (Appendix E & F). The juvenile cohorts did not deviate from the expected 1:1 sex ratio, whereas the 2011 adults displayed sex ratio of 1:5 with male biased overrepresentation (Table 2).

Table 2. The genotypic sex ratios of spawning adults of 2011 and yearly sampling groups of juvenile cohort of 2012. * marks a significant deviation from the expected 1:1 sex ratio in the 2011 adults (Fisher's exact test, P-value 1.144e-06). Both sexes were present in the expected 1:1 ratio in the juveniles.

	+0y 2012	+1y 2013	+2y 2014	+3y 2015	Adults 2011
Male	394	448	91	33	83
Female	395	447	91	39	17
In total	789	895	182	72	100
Sex ratio	1:1	1:1	1:1	6:5	1:5*

3.1.2. Sea-age candidate gene loci genotyping success

The *vgll3$_{TOP}$* locus was successfully genotyped in 1643 of the total 2096 individuals (including additional adults from the same cohort year). Of these 1643 individuals with known *vgll3* genotype, 1627 were also successfully assigned to either sex, allowing testing for selection between the sexes (Table 3). As expected, there were no significant deviation in genotyping success of these *vgll3* locus between the sexes. There were some variation in the genotyping success of the *vgll3$_{TOP}$* locus between the year groups, but this was not due to qualitative differences between these groups but rather due to batch effects during library preparation and sequencing. Likewise, the relatively low genotyping success of the +2y 2014 was due to unintentional exclusion of the primers targeting the *vgll3$_{TOP}$* locus from the primer pool prepared for the 3rd sequencing library. The exact nature of this error is not known, but it's most likely due to a pipetting error. On the other hand, the particularly good genotyping success of the +3y 2015 and the Adults 2011 resulted from the re-sequencing of the 6th library that contained all of these samples. The extra data provided by this re-sequencing merged together with the previous data from the first sequencing of this library increased the overall coverage on these samples allowing better genotyping success with lower error rate (Appendix I).

Table 3. *vgll3TOP* locus genotyping success in each sample group. The +0y 2012 and +1y 2015 represent the average genotyping of the *vgll3* marker locus. The low genotyping success of the +2y 2014 year group was due to unintentional exclusion of the primers targeting the *vgll3TOP* from the primer pool prepared for the 3rd library that contained 96 samples of this group. The particularly good genotyping success of the +3y 2015 and the Adult 2011 groups was result from the re-sequencing of the 6th library that contained all samples of these groups. The extra sequence data merged together with the original one allowed more accurate genotyping of these samples.

	+0y 2012	**+1y 2013**	**+2y 2014**	**+3y 2015**	**Adults 2011**
Male	330	349	45	33	83
Female	332	356	43	39	17
In total	662	705	88	72	100
Genotyping success	80.1 %	77.4 %	47.8 %	96.0 %	100 %

The *six6* locus was successfully scored in almost all individuals included in the final dataset (2056 individuals out of the 2059). Thus, there were no differences in genotyping success between the cohorts nor sexes.

The *akap11* genotype was genotyped in 1783 individuals of the 2059 in the final dataset. However, the overall distribution of successful genotyping among cohorts vary among age groups batch dependently, since the amplification of the locus in the 5th Ion Torrent library failed completely, resulting in complete absence of genotypic information in the +3y 2015 cohort and in the 54 adults from 2011 that were part of the original dataset and had parentages assigned to 2012 offspring. It was later shown that this failure was due to the usage of incorrect primer pair in the initial amplification of the locus in the PCR-1. The primers used instead were earlier configurations of the primers targeting the *akap11* locus that turned out non-functional.

3.2 Sexual selection targeting the sea-age loci in spawning Atlantic salmon adults

The hypotheses exploring the occurrence of sexual selection in Atlantic salmon associated with the *vgll3* locus and other candidate sea-age gene loci were tested at two different levels. First, the trans-generational genotype frequencies were studied in order to detect if certain *vgll3* genotypes were enriched in the offspring, signalling of better breeding success conveyed by certain genotype in the adults. Second, the genotypes of the known mating pairs were studied to establish if assortative mating was present among the spawning adults based on the *vgll3* genotypes. In the context of the latter approach,

segregation distortion was also studied by observing the ratios of offspring of certain genotype born for the parents of known mating pairs to assess if further post-copulatory mechanisms work in favour of certain *vgll3* alleles during the fertilization.

3.2.1 Trans-generational genotype frequencies and *vgll3* related breeding success

The generational difference in the sea-age genotypes were studied in order assess whether certain sea-age genotypes result in better breeding success. First the *vgll3* related differences in the reproductive success were studied by comparing the observed offspring genotype frequencies to the expected offspring genotype frequencies in order to assess if some genotype was enriched in offspring. The expected offspring genotype frequencies were acquired by calculating the likelihood of a certain allele from the female side to coincide with certain allele from the male side at the fertilization. These likelihoods were then translated to genotype counts by multiplying them with the total count of the 662 +0y 2012 juveniles with known *vgll3* genotypes in the dataset. The expected offspring genotype frequencies were not simply derived from Hardy-Weinberg equilibrium, since such assumptions as equal sex rations and allele frequencies between the sexes were not met here. This difference was highly significant (Fisher's exact test, P-value < 0.001) (Table 4). The *LL*-genotype was considerably more common than expected (+ 11.8 %), whereas both the heterozygous and the *EE*-genotype showed modestly scarcer than expected (- 7.8 % and - 4.1%, respectively).

Table 4. The difference between the expected and observed offspring *vgll3* genotype frequencies. This difference was highly significant (P-value < 0.001).

Genotype	Expected offspring count	Observed offspring count	Difference (%)
LL	186	264	+ 11.8
LE	344	293	– 7.8
EE	132	105	– 4.1

The trans-generational change in the genotype frequencies was further studied by comparing the observed genotype frequencies between two generations. The adult genotype frequencies were compared to the +0y juveniles as whole and between the

sexes. When the adult males were compared to the juveniles, the difference in the genotype frequencies at the *vgll3* loci (vg*ll3*$_{TOP}$ locus and the two other *vgll3* associated sea-age peak loci) and the *six6*$_{TOP}$ locus were the most significant of all loci studied exceeding alpha threshold of 0.01 (Figure 6). The two other markers on the *vgll3* peak region also exceeded the 0.001 threshold. In both of these alpha thresholds genomic control λ was factored in to account for genomic inflation. When the adult female and the +0y juvenile genotype frequencies were compered, none of the sea-age associated loci exceeded similar thresholds nor ranked among the most significantly deviating loci. On the contrary, the *vgll3*$_{TOP}$ genotypes were virtually the same (P-value = 1). However, the deviation in the *six6*$_{TOP}$ genotype frequencies exceeded the alpha threshold of 0.05 adjusted by genomic control λ. Several other markers were still more significantly deviated. Overall, there were more significant deviations in the genotype frequencies between the adult males and juveniles than the adult females and juveniles as well as more genetic inflation (Figure 6).

As significant results are expected by chance alone (i.e. false positives) which may be further elevated by genomic inflation, Q-Q plots were constructed to further observe the nature of the significance of deviations in the genotype frequencies between the +0y juveniles and adult males and females and the effects of genomic inflation (Figure 7). Between the adult males and juveniles the deviation in the vg*ll3*$_{TOP}$ locus and the two other *vgll3* associated sea-age peak loci and the *six6*$_{TOP}$ locus appeared more significant than the expected most significant values in the null distribution, seen as separation of these loci from the expected regression (Figure 7a). However, of these loci only the vg*ll3*$_{TOP}$ and one of the two other *vgll3* peak region markers did not fall within the confidence intervals. Between the adult females and juveniles the several most highly deviating loci expressed slightly higher significance than expected in the null distribution, *six6*$_{TOP}$ among these (Figure 7b). However, all of these fell well within the confidence intervals.

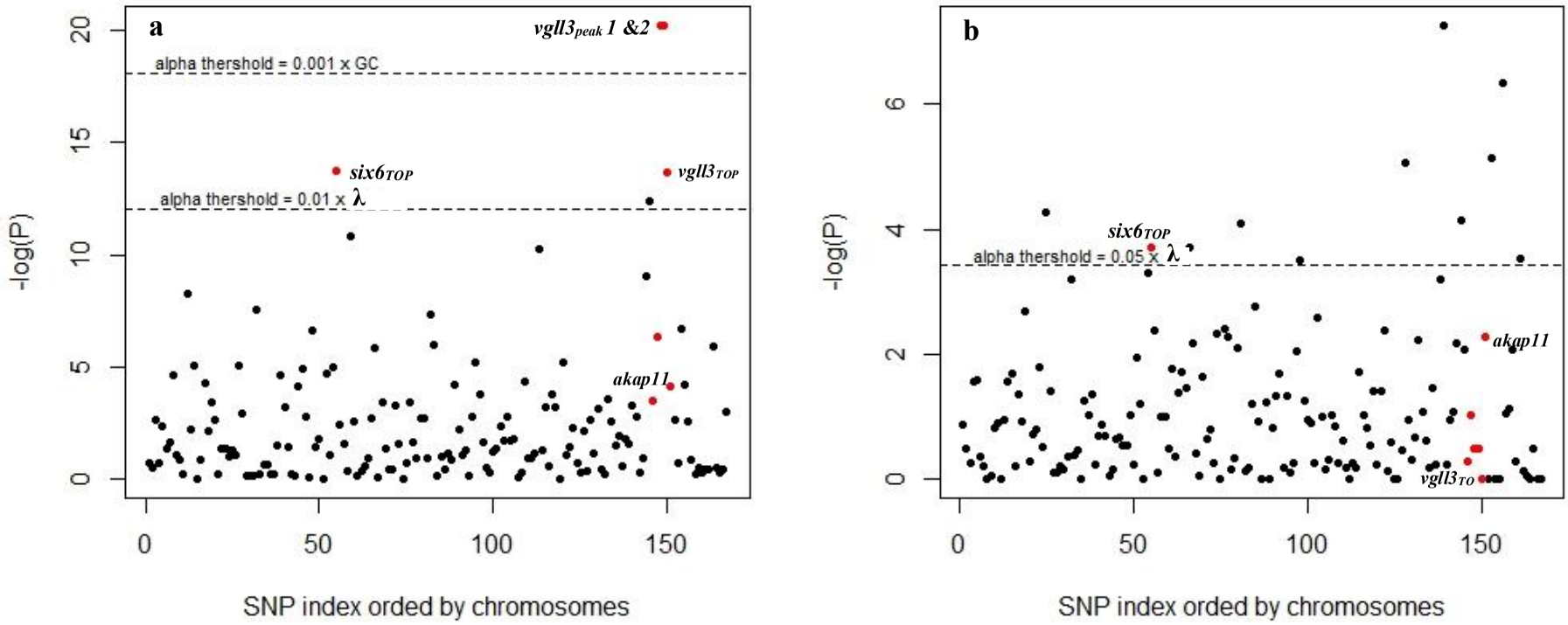

Figure 6. Significance of genotype frequency deviations between adult and +0y juvenile Atlantic salmon (Fisher's exact test, -log scale). a, Adult male genotype frequencies and (**b**) adult female genotype frequencies compared to the juvenile genotype frequencies. The SNPs are ordered by the chromosome and the chromosomal position they are located at, and the sea-age associated SNPs are coloured in red. The deviations were greater between adult males and juveniles (**a**), and alpha thresholds 0.01 and 0.001 adjusted by genomic control λ (observed median P-value deviated by median of chi-square distribution with two degrees of freedom) were applied. Above the first threshold ranked *vgll3*$_{top}$ and two other *vgll3* associated loci, *six6*$_{TOP}$ and one other marker that has no association to the sea-age, and above the latter two *vgll3* associated markers but not the *vgll3*$_{TOP}$. The deviations in genotype frequencies between adult females and juveniles (**b**) were less drastic, and similar thresholds couldn't be applied as in comparison of adult males and juveniles. Instead, alpha threshold 0.05 adjusted by λ was applied, and the only sea-age associated locus that ranked above this threshold was *six6*$_{TOP}$ together with several other markers.

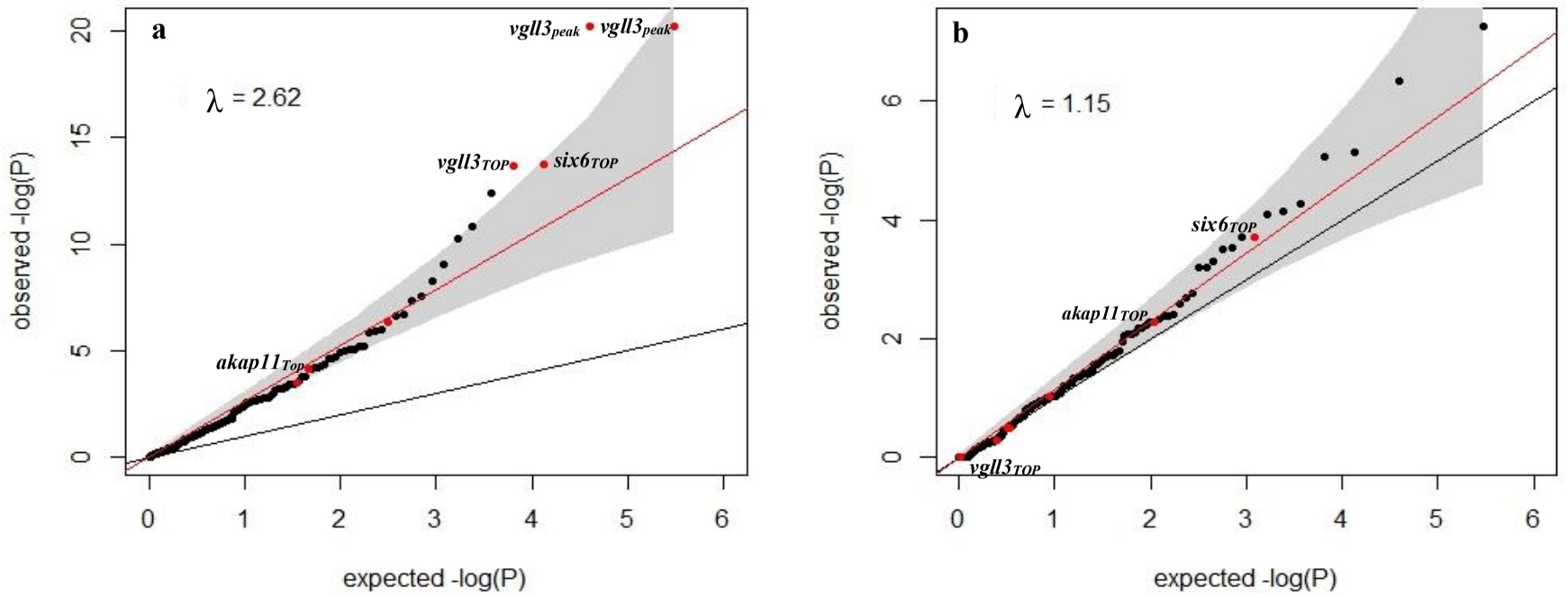

Figure 7. Chi-square distribution quantiles with two degrees of freedom of Fisher's exact test -log P-values yielded from comparison of genotype frequencies on 167 loci of two subsequent generations of Atlantic salmon plotted against random chi-square distribution with two degrees of freedom. The black line represent the expected alignment of quantiles of two similar distributions and the red line represent similar expectation after correction for genomic control λ for genetic inflation (observed median P-value deviated by median of chi-square distribution with two degrees of freedom). The grey shading indicates a 95 % confidence interval. **a,** a Q-Q plot constructed based on the P-values resulting from comparison of the adult male and +0y juvenile genotype frequencies. After correcting for genomic inflation, the quantiles aligned as expected. The *vgll3*$_{TOP}$ and two other *vgll3* peak region associated markers and the *six6*$_{TOP}$ are in the most extreme quantiles and separate as outliers indicating higher significance than expected by random. However, only *vgll3*$_{TOP}$ and one of the two other *vgll3* peak region markers separate from the confidence intervals together with one other marker from lower quantile. **b,** a Q-Q plot from comparison of genotype frequencies between adult females and +0y juveniles. After correcting for genomic inflation, the quantiles aligned as expected. Some separation among the markers from the expected was observed in the most extreme quantiles, and among these were the *six6*$_{TOP}$. None of these markers however separated from the confidence intervals. Overall, there were less deviations and genetic inflation between the adult females and juveniles than the adult males and the juveniles.

3.3.2 Non-random pairing

Based on the pedigree established by Ellmén (2015), information about past mating pairs was available. Overall there were 20 adults with known mating partner that formed 14 known mating pairs (Appendix B.). Six of these adults were females, of which three were homozygous for *LL,* and three were heterozygote (*LE*) on the $vgll3_{TOP}$ locus, and 14 were males, of which four, seven and three were genotyped as *LL, LE* and *EE*, respectively. The values in the contingency table tested were the occurrences of individual of certain genotype to mate with another individual of certain genotype. No assortative mating was detected among these spawning Atlantic salmon adults based on their $vgll3_{TOP}$ genotypes (Fisher's exact test, P-value = 0.24).

Segregation distortion, i.e. outcome where genotype frequencies of offspring of specific mating pairs are not according to Mendelian expectations possibly due to viability selection, was tested to explore sexual selection hypothesis with a wider scope. It was speculated that *vgll3* genotype ratios may deviate from expected Mendelian ratio due to differences in genetic compatibility, fertilization success or early survival between the genotypes. This was tested by comparing the offspring genotype ratios born for known mating pairs where one or both parents were *LE* heterozygotes to those expected under Mendelian segregation. However, this testing was severely limited due to the small sample size. In total there were 14 offspring with known *vgll3* genotypes born for mating pairs where both parents were heterozygous. 5, 3 and 6 of these were *LL, LE* and *EE* genotypes, respectively. When this ratio was compared to expected Mendelian ratio of 1:2:1, no statistical significance was observed (Fisher's exact test, P-value = 0.44, Appendix C). In the second setting, where offspring that were born for mating pairs where one parent was *LL*-homozygous genotype and the second heterozygous *LE* genotype were studied, only total of 10 offspring with known *vgll3* genotype were present in the dataset. All of these offspring individuals expect one were heterozygous, but for such a small sample size statistical testing wasn't truly meaningful (Fisher's exact test, P-value = 0.14, Appendix C). Thus, no segregation distortion based on the *vgll3* alleles carried by the gametes was detected in Atlantic salmon from this dataset.

3.3. Testing for selection on vgll3 genotypes during the juvenile fresh water phase

In order to test of natural selection differentially act on *vgll3* genotypes at the freshwater phase, we compared the change in genotype frequencies within a cohort (i.e. 2012 hatched juveniles from the lower Utsjoki) along four age groups (from 0+ to 3+, Appendix D.). The *vgll3* genotype frequencies were also compared within a cohort year between the sexes. Results were not significant in any occasion (Appendix D), thus, we found no evidence of *vgll3* genotype related selection, migrating behavior differences or differences in maturation during the early juvenile years in Atlantic salmon.

4. DISCUSSION

In this study, I investigated whether genetic variation associated with the sea-age life history trait is associated with fitness differences at different stages in the life history of Atlantic salmon. This was done both by studying genotype dependent breeding success in the context of sexual selection and by studying the genotype dependent survival of individuals at early life history stages.

4.1 Sexual selection targeting the sea-age genotypes

4.1.1 Trans-generational sea-age genotype frequencies

The sea-age genotype frequencies were compared between adults and juveniles in order to assess whether adults with certain genotypes succeed better during the breeding and/or if there is some drive for production of progeny of a certain genotype. In contrast to the null hypothesis, there was a significant deviation in *vgll3* genotype frequencies observed between adults and +0 individuals. There was an excess *LL*-genotype offspring (+ 11.8 %) while both the heterozygous *EL* and homozygous *EE*-genotypes were lower than expected in the +0 offspring cohort compared to the adult 2011 cohort (- 7.8 % and - 4.1 %, respectively). This result most likely reflects the better breeding success of larger Atlantic salmon. Such individuals are more likely to possess at least one *L* allele (Barson *et al.* 2015) and more sizeable females also have more eggs and can occupy the best

nesting sites and produce more offspring with more resources to start with. Further, bigger males are preferred by the females and able to drive away the competing smaller males (Fleming 1996; Aas *et al.* 2011; Mobley *et al.* 2019).Thus, the *L*-alleles that ultimately strongly contribute to this phenotype of bigger size are more likely to be transmitted to the next generation of Atlantic salmon. It could also be speculated that mate choice or non-random allele segregation mechanism favoring the *L*-allele or the *LL*-genotype would exist, but this seems unlikely as there's no evidence that it would be mutually beneficial for the sexes, since the most sizeable anadromous phenotype is not unequivocally the most advantageous phenotype in the male Atlantic salmon (Fleming 1996).

The male and female adult and the offspring *vgll3* genotypes frequencies were also compared. A significant difference was observed between the adult male and the offspring genotype frequencies (P-value < 0.001), whereas there was no such difference detected between adult females and the offspring (P-value = 1). The deviation between the adult male and the offspring genotype frequencies remained significant after the correction for genetic inflation (P-value < 0.01) (Figure 6). These results indicate that the offspring *vgll3* genotype frequencies resemble more the female than the male genotypes. Altogether, these results appear coherent with the better success of the bigger fish during the spawning and suggests the role of sexual selection acting on males: most, if not all, females manage to produce relatively large numbers of offspring, whereas due to the female choice and the male-male competition the bigger male fish tend to have better breeding success and the smaller 1SW males have more minor contribution to the spawning (Appendix E) (Järvi 1990; Fleming 1996; Mobley *et al.* 2019). Thus, it can be that the *L*-alleles are more likely to be transmitted to the next generation of Atlantic salmon, as the more sizeable individuals with better breeding success are more likely to carry this allele. The *E* allele most likely persists in the population as the associated 1SW life history allows sufficient breeding success in males.

The genotype frequencies of the second sea-age candidate region on chromosome 9 harboring *six6* were also observed to be significantly different between the +0y juveniles and the adults in both males and females, however, as for *vgll3*, the deviation was much stronger in males (P-value < 0.01 and <0.05, respectively) (Figures 6 & 7). The *six6* locus has been shown to be strongly correlated to population structure in Atlantic salmon with very large differences in allele frequencies being reported also in the Teno river system (Barson *et al.* 2015; Pritchard *et al.* 2018). Barson *et al.* (2015) also showed that the

frequencies of the *L* alleles on the $vgll3_{TOP}$ and $six6_{TOP}$ locus tend to correlate within Atlantic salmon populations, but such association was not evident here. On the contrary, whereas the *vgll3 L* allele was more abundant in the lower Utsjoki region population studied here, the *six6* locus was almost fixed for the *E* allele. Furthermore, the *E* allele of the *six6* locus was enriched in offspring in a similar manner to the *vgll3 L* allele when the expected and observed offspring genotype frequencies were compared (Fisher's exact test, P-value < 0.001). It remains elusive why there is such deviation in the genotype frequencies on this locus present between the two generations studied here. Pritchard *et al.* (2018) showed evidence of local adaptation on the *six6* locus, and the causative natural selection was speculated to target traits that are known to be associated or regulated by the *six6* gene in other vertebrate species, such as eye and retinal development and hypothalamic functions e.g. circadian rhythm and gonadotropin-releasing hormone production (Larder *et al.* 2011; Conte *et al.* 2010; Watanabe *et al.* 2012). These or other unknown *six6* associated traits could affect the survival at sea and cause the observed deviation in the genotype frequencies. Furthermore, the alleles that are possible less beneficial at sea (affect survival negatively) could be favored by selection in freshwater environment or during the spawning in sexual selection, analogically to the *vgll3 L* allele in male Atlantic salmon.

Genotype frequencies of the *akap11* appeared significant between the adult males and +0y juveniles at first (P-value = 0.016) but after correcting for genomic inflation this deviation didn't exceed any alpha threshold (Figure 6) and settled in the expected null distribution (Figure 7). Between the adult females and juveniles there wasn't any significant deviation even before correction for genomic inflation (P-value = 0.1, Figure 6), and similarly as in the case of adult males, this deviation followed the expected null distribution (Figure 7). These results indicate that the different *akap11* genotypes don't result in differences in breeding success or survival. Would the *akap11* be a major contributor in the sea-age associated peak region on the chromosome 25 similarly to the *vgll3*, similar patterns could have also been expected in the trans-generational genotype frequencies as discussed above in the case of the *vgll3*. Besides the effects on the sea-age, the different *akap11* genotypes could have possibly altered the sperm quality of the adult males due to its known role in the spermatogenesis and sperm functioning and hence the breeding success via sperm competition (Reinton *et al.* 2000), but as stated before, this appears unlikely as no truly significant differences in the genotype frequencies were

observed. The modest level of significance observed before applying genomic control was likely rather due to the linkage disequilibrium between the *akap11* to the *vgll3* locus than the actual effects of the different *akap11* genotypes. However, due to the absence of the genotypic information on the *akap11* locus of the 54 adults that were part of the original sample material, the sample size for this specific locus is relatively small and more data could provide a better insight on the matter and adduce some trend that was not detected here.

The potential contribution of the precocious parr to the +0y 2012 offspring genotype frequencies also remain elusive. Since the sires were assigned only for 177 of the 803 +0y 2012 offspring (22 %) in this dataset, the identity of most sires remain unknown. These unknown sires can either be other anadromous males that have migrated back to Utsjoki to spawn after the sea period and weren't captured by Ellmén (2015), or they are precocious male parr that have matured before the smoltification and the migration to the sea. As there is no information of the missing sires, the overall ratio of the anadromous sires to the precocious male parr sires remains also unknown. However, previous studies suggest that the proportion of the mature parrs of all males in Utsjoki is 25 % at highest, probably somewhat less (Heinimaa and Erkinaro 2004). Further, the breeding success of individual mature parr has been shown to be low compared to anadromous males, but that the overall mature parr contribution in offspring fertilization can be up to 40 % (Thomaz, Beall, and Burke 1997; Tentelier *et al.* 2016). As most of the sires weren't assigned for the Utsjoki +0y 2012 juveniles, some contribution by the mature parrs seems plausible here also.

Besides the impact size of the mature parr contribution to the offspring genotypes, the quality of the contribution remains similarly elusive, since there is no knowledge of the genotypes that the mature parr carry and whether the precocious maturation is related to the genes studied here or if they are rather due to other genetic and environmental factors. The early maturation has been shown to relate to faster growth rate in the natal river (Simpson 1992; Saunders, Henderson, and Glebe 1982). Due to the faster growth, these individuals have also more resources available to allocate for gonadal development and gamete production. The faster growth rate is often evident from early on after the hatching implying that there is a genetic component underlying that physiological trait (Aubin-Horth and Dodson 2004). Some evidence also exist that the precocious parr maturation is related to the *vgll3* locus so that the *E* allele is associated to the faster growth as parr and

thus to precocious parr phenotype (Lepais *et al.* 2017). However, this evidence is incomplete but may suggest that the precocious parr phenotype could elevate the overall fitness of the *EE* males and alleviate the negative effect of the trade-off in survival and breeding success, and hence facilitate the persistence of this allele in the populations. Furthermore, if the precocious parr phenotype is indeed coupled with the *E* allele and would these individuals contribute in any major way to the spawning, enrichment of the *E* allele in the offspring genotypes could be expected. Since the opposite was shown to be true in this study setting, it can be speculated that either the precocious parrs didn't contribute to the spawning in any major way or the effect of *E* allele of the *vgll3* locus needs further dissection.

4.1.2 Mate choice related breeding success

Mate choice related sexual selection in Atlantic salmon was studied by examining the known mating pairs in the study material. My hypothesis was that the suggested resolution of sexual conflict via sex dependent dominance should lead to heterozygous advantage, thus resulting in mate choice for ensuring good quality genes as *vgll3* heterozygosity in offspring. Earlier examples of mate choice for "good genes" in Atlantic salmon due to heterozygous advantage have been documented for MHC (major histocompatibility complex) coding genes (Evans *et al.* 2012; Landry *et al.* 2001). MHC are immunodefence related receptors used for pathogen recognition, thus the wider variety of these receptors provided by heterozygosity leads to better pathogen recognition and elevated fitness in vertebrata (Penn, Damjanovich, and Potts 2002; Neff and Pitcher 2005). However, no *vgll3* related mate choice was observed among the studied spawning Atlantic salmon. One straightforward explanation supported by this observation is that that there isn't any fitness advantage to gain would the females favor the production of heterozygous offspring. However, the lack of assortative mating based on *vgll3* genotypes could also be due there not being a direct mechanism by which the females (the choosing sex in the Atlantic salmon) could infer the underlying genotype in males. For example, another Salmonidae species, Archtic charr (*Salvelinus alpinus*), has shown to be able to discriminate between different MHC genotypes in other individuals based on the olfactory cues (Olse 1998). Human olfactory receptor genes are also shown to be associated with MHC genes, indicating that behavioral preference for certain MHC

haplotypes in humans may also be olfactory mediated (Fan *et al.* 1995). If no similar cue exists for conveying *vgll3* genotype at the phenotypic level, assortative mating has no mechanism to target the *vgll3* genotypes. It could be argued that the sea-age genotype is evident at the phenotypic level as the size of the fish, but this wouldn't necessarily allow reliable distinction between the homo- and heterozygous individuals in males, as the *E* allele is dominant (Barson *et al.* 2015). Moreover, as *LL*-homozygosity is more common in the females, the *EE*-males could be preferred in order to favor production of *vgll3* heterozygous progeny. However, the bigger males that are often *LL*-genotype, tend to dominate the spawning grounds and they are primarily favored by the females. This would suggest that female choice is based primarily on the size of the male and the ability of the male to establish dominance on of the spawning grounds as shown before.

The overall plausibility for such mechanism to exist that would allow the Atlantic salmon individuals to discern the *vgll3* genotypes of other individuals can be speculated. Often mate choice targets signals of overall quality of the mate, and selection targeting specific genotype is more rare (Thibert-Plante and Gavrilets 2013; Tregenza and Wedell 2000). Moreover, while certain mating partners are preferred by all individuals of the opposite sex when mate choice targets signals based on the overall condition, preferences for genetic compatibility drives individuals to prefer different mating partners (Tregenza and Wedell 2000). This kind of assortative mating system targeting certain genotypes of one locus is this far only well documented for vertebrate MHC haplotypes, which are of ancient evolutionary origin and has been under strong selection pressure for millions of years since the common ancestor of all vertebrate (Piertney and Oliver 2006). In this context, highly elaborate selection mechanisms targeting these genes are plausible, whereas there is no evidence that the *vgll3* and its sex dependent dominance expression mechanism are equally ancient origin that have similarly been under constant selection pressure. Hence, it's less likely that similar mechanisms would target this gene, even if the fitness gains for mate choice based on the *vgll3* compatibility over the MHC genes would be higher. Likewise, as there already exist MHC haplotype based mate choice for compatible genes, similar mechanism to exist and function in parallel is unlikely (Thibert-Plante and Gavrilets 2013). Further, even though assortative mating is known to exist in spawning Atlantic salmon, the male-male competition reduces the female possibility for mate choice (Fleming 1996). In these circumstances where sexual selection is already targeting multiple traits in Atlantic salmon and the mate choice is somewhat limited, the

evolution of a new additional target for sexual selection is expected to be constrained. Hence, the existence of such mechanism for mate choice based on the *vgll3* genotype isn't necessarily the most likely scenario, but nevertheless possible as similar mechanism exists for the MHC genes. Further studies with larger dataset could help to shed light on this matter.

Besides the assortative mate choice, the segregation of the *vgll3* alleles were inspected for detecting possible segregation distortions. However, no significant deviations from the *vgll3* related Mendelian offspring ratios were detected in the offspring broods produced by pairs of parent fish that were both heterozygous and pairs where one parent was *LL*-genotype and the other heterozygous (Appendix C). This result would suggest that there is no *vgll3* genotype related post-copulatory mechanism that promotes heterozygosity in the offspring. However, the power of this analysis was notably weak due to the small number of known mating pairs suitable for this testing and the offspring assigned for these pairs. Therefore, as in the case of the individual mate choice discussed above, further study is needed before reaching a conclusion. Like in the case of the pre-copulatory mate choice discussed before, the MHC genes again have been shown to function in the post-copulatory selection in the Atlantic salmon (Yeates *et al.* 2009). Despite that the effect was inversed on the gamete level compared to the mate choice so that the genetic similarity was preferred instead heterogeneity, there were again discrimination in the fertilization success based on this one locus, providing a model how a similar mechanism could act on the *vgll3* locus.

4.2 Sea-age genotype related survival

4.2.1 Sea-age genotype related juvenile survival in natal freshwater river

No significant differences in the *vgll3, six6 or akap11* genotype frequencies were detected in the 2012 hatched juvenile cohort between annual samplings or between the sexes on different years (Appendix D). This result would strongly imply that there are no clear sea-age genotype dependent differences in the offspring physiology, or behavior that would affect the survival or abundance of individuals during the fresh water phase. This was assumed plausible especially as the *vgll3* gene has been shown to function as adiposity

regulator (Halperin *et al.* 2013), which is a physiological process tightly linked to the growth rate, body size, fat content and overall physiology (Simpson 1992). Further, precocious maturation have been linked to the *vgll3* region, and this trait has likewise been associated to the growth rate and fat content (Lepais *et al.* 2017). Thus, the different *vgll3* genotypes could result in differences in the size and body content of Atlantic salmon parr. Different sized parrs again are known to utilize their environment somewhat differently by feeding with different pray and habit different sites of their natal river (Heggenes 1990; Keeley and Grant 2011). Moreover, besides that the faster growing juveniles are more likely to mature as parr, the bigger juveniles have been shown to go through the smoltification process earlier (Saunders, Henderson, and Glebe 1982). Finally, the *vgll3* gene could be associated to these or other behavioral and physiological traits through other unknown processes. However, since no differences in the *vgll3* genotype frequencies were observed between the age groups or the sexes, it seems that the *vgll3* genotype doesn't differentiate the juvenile Atlantic salmon in any major way. Lack of power to detect an effect seems unlikely as the sample size was relatively high for each annual sample. Also notable in this data was the evenness of the sex ratios (Table 2) which contrasts the 1:5 female:male ration of the breeding adults. This result indicates that there are no sex related differences in the survival or dispersal of Atlantic salmon parrs during the first three years of their freshwater period in their natal river. The most common smoltification age in Teno salmon is 4 or 5 years (Erkinaro *et al.*, 2018), but samples from older parr were not available for this study so it is not possible to conclude that this conclusion holds for the entire freshwater period, but strong selection in fresh water seems unlikely at least to three years of age.

4.2.2 *vgll3*-related effects on marine survival

Even though not directly studied here, the comparison of the offspring and adult sex ratios and *vgll3* genotype frequencies can yield some information and allow speculation about *vgll3* related survival during the growth period at sea. The most striking difference between these demographic groups is that while the Atlantic salmon still reside their natal freshwater river before smoltification and migration to sea, the sex ratios are particularly equal, whereas upon the spawning when the adults migrate back to their natal sites after the sea period, there are five times the males compared to the females. All of the females

expect one were multi sea winter fish, and 80 % of the males were 1SW fish (Appendix E). There is simultaneously significant difference in the *vgll3* genotype frequencies between the sexes so that the *LL*-genotype is more common in females whereas the heterozygous and the *EE*-genotype are present in large numbers in males (Fisher's exact test, P-value = 0.03). These results fit the previously established understanding about the increased mortality risk of Atlantic salmon together with a life history of multiple sea winters, and how the *vgll3* genotype is related to different utilization of the life histories between the sexes (Hansen and Quinn 1998; Barson *et al.* 2015). The excess of males can be explained by the more frequent early maturation at sea and that minimizes the risks of early mortality, and the observed *vgll3* genotypes are such that likely result in the expression of this life history (Barson *et al.* 2015). Conversely, the females are present in lesser numbers since they spend consistently multiple winters at sea and thus perish more often before the spawning. Like the male genotypes, the female *vgll3* genotypes and expected phenotypic patterns reflect the observed life histories (Appendix E). Both the heterozygous and the *LL*-genotype that result mainly in the MSW phenotype were present in equal numbers, while the *EE*-genotype was virtually absent. However, there is known to exist lot of variation in the sex ratios of returning Atlantic salmon between the different populations (Fleming, 1998; Erkinaro *et al.*, 2018). Hence, these patterns in the sex ratios and the *vgll3* genotype frequencies are not necessary similar in other populations. Moreover, a lot of annual variation is known to exist in the Atlantic salmon breeding system (Fleming, 1998; Erkinaro *et al.*, 2018) Events both in the freshwater and sea such as epidemics or alterations and anomalies in the temperature, food supplies, and water flowrates can introduce temporal variation in the amount of annual spawners and bias the sex ratios, as the sexes are subjected to these events somewhat differently due to the differences in the utilization of the life history strategies. Thus, overall the speculations based on the sex ratios and genotype frequencies of one adult and one juvenile cohort in a single population may not be directly generalizable to all populations of Atlantic salmon.

The near complete absence of the females of the *vgll3 EE* genotype is somewhat peculiar (Appendix E). There were only two females of this genotype present, and of these only one were 1SW phenotype. Whereas the male *vgll3* genotypes reflect the expected and observed life histories, there is no obvious explanation for the small amount of *EE* females. Even though this genotype results in longer sea period in females than males,

these females are expected to mature earlier on average and should have thus higher relative survival (Barson *et al.* 2015). If the juveniles hatched on 2012 are used as an approximation from the genotype frequencies of the spawning adult cohort of 2011 before their sea-migration, higher *EE* genotype frequency could be expected due to elevated survival despite that genotype is present in lesser number among the offspring. It could be that due to the nest competition between females the smaller females that are *EE* genotype are forced to spawn on less optimal locations and were thus further from the capturing site (Fleming 1996). However, there is no evidence to support this, and tendency of females for later maturation even if *EE* genotype together with assumed lower genotype frequency to start with could maybe still lead to such numbers as observed in this study especially if factoring in any chance or small capture bias.

Some bias in the adult sex ratios and *vgll3* genotype frequencies of this data can indeed result from some kind of capture bias. For example, Ellmén (2015) speculated that the kype, a specialized hook structure and a secondary sexual character of male Atlantic salmon (Fleming 1996), would facilitate the male capture and result in over representation of males in the final catch. However, this was thought to be unlikely since the females stick nearby to the spawning grounds and are thus likely to be caught at some point (Aas *et al.* 2011; Ellmén 2015). Whether or not the kype affected the catch, some other unaccounted factor may have still introduced some bias to the capture rates between the sexes. While studying long term trends in the maturation times in Teno river Atlantic salmon, Erkinaro *et al.* (2018) noted that different fishing methods had yielded different amounts of fish of different maturation ages. However, given that Mobley *et al.* (2019) report a similar male-biased sex ratio in four additional sampling years from the same study site, it seems likely that the observed sex ratio is accurate.

5. CONCLUSIONS

In 2015, Barson *et al.* identified three candidate genes *vgll3, six6* and *akap11* in two genomic regions affecting the age at maturity in Atlantic salmon. Furthermore, a sex dependent dominance effect was observed in genomic region with strongest association to sea-age in close vicinity of the candidate gene *vgll3.* This was assumed to be an adaptation to sexual conflict allowing the expression of more optimal life history

phenotype in both sexes. However, the the actual effects of the different genotypes of sea-age associated genes on fitness were left un-explored. In this study, my objective was to fill some of these knowledge gaps by studying both the breeding success and survival of individuals of different genotypes having the ephasis on the *vgll3* gene due to its strongest association to the sea-age and sex specific effects.

This study of sea-age candidate gene related differences in breeding success revealed a significant difference in genotype frequencies of *vgll3* and *six6* between the two generations of Atlantic salmon studied here. *L* allele of the *vgll3* gene was found to be enriched in juveniles, and it was reckoned to be due to the better breeding success of the bigger male that are more often carriers of *L* alleles (Fleming 1996; Barson *et al.* 2015). However, the majority of adult males were heterozygous or *EE*-genotype, since these genotypes result in shorer marine period and in better survival. In case of the candidate gene *six6*, the *E* allele was found to be enriched in juveniles, and there was no immediate satisfacory explanation for this trend. Furthermore, this was somewhat contradicting the findings of Barson *et al.* (2015), as they found a correlation between frequencies of alleles of *vgll3* and *six6* with similar phenotypic effect on sea-age. No significant differences in *akap11* genotype frequencies were found beween the two generations and thus it's unlikely that this gene has any effect on breeding success or survival. The role of the sexually mature precocious parrs may somewhat obscure these results, however, as there is evidence that suggests they have relatively minor contribution to the next generation and an opposing effect than the observed overall trend between sea-age genotypes and breeding success (Heinimaa and Erkinaro 2004; Lepais *et al.* 2017). *vgll3* related mate choice was also studied, as well as possible postcopulatory selction causing segregation distortion between the alleles. No *vgll3* related effect was detected. However, these tests were severly limited due to the scarce information of known mating pairs and these results remains inconclusive. Adressing the second hypothesis of this study, no sea-age candidate gene related effect to the juvenile survival was observed. Thus, even though these genes evidently induce or participate to major physiological changes in later Atlantic salmon life history, they don't appear to have similar role on pre-smoltification juvenile life history stage.

6. FUTURE RESEARCH

Both the breeding success in context of sexual selection and survival of Atlantic salmon in relation to their sea-age candidate gene genotypes were studied here. However, the breeding success was studied only by observing the change in genotype frequencies over one generation gap and by observing the mate choice within one generation of breeding Atlantic salmon. Even though this approach is adequate to discern trends in genotype related effects on breeding success, it cannot yield any information about fluctuations and stablility of these trends. Thus, extending the study conducted here over multiple generations would provide a broader and more precise view over the matter. Furthermore, as Atlantic salmon has its range in whole northern hemisphere, inclusion of multiple different populations would likewise add to the uderstanding of how the relationship between sea-age genotypes and breeding success possible varies between populations and what are the possible environmental factors contibuting to these outcomes, e.g. could the breeding success of smaller males be higher in some populations and the *E* rather than the *L* allele as found here be enriched in the following generations. The whole aim to assess the effects of the sea-age genotypes to the breeding success was somewhat limited due to the small sample size, especially since there were very few adult females and known mating pairs in the dataset. Hence, for future studies, capture of more adults should be prioritized, especially females.

While adressing the second aim of the study about the effect of the sea-age candidate gene genotypes to the juvenile survival, similarly as with the studies about the genotypic effect on breeding success, only one cohort at a single location was studied. Hence, in order to generalize the findings here, study of multiple generations from several populations would be beneficial. The monitoring period should also be extended over the most common smoltification ages of juveniles, as this may be affected by the sea-age candidate genes and result in changes in the genotype frequencies in those older juvenile age groups that were not included to this study. In contrary to the studies of breeding success that were partly limited due to small sample size, the sample size wasn't a limiting factor while the sea-age candidate gene associated juvenile survival was studied and additional samples would have been unlikely to reveal any new findings. However, the post-smoltification survival coud bring new insights of the effects of the sea-age candidate genes on the survival of Atlantic salmon. For example, the different genotypes

could affect the physiology or behaviour of the adult fish during the marine phase causing differences in survival other than simply prolonging the vulnerable marine period or affect the timing of smoltifiation resulting in size differences among Atlantic salmon smolts migrating to sea altering the susceptibility to predation.

In the light of current study, some streamlinings and improvements to the workflow may be suggested for future research. Here, there was a continuos optimization process going on, and even though this simplified the protocol and hastened the preparations of an individual library, the process on itself was time consuming and complicated the data gathering. Thus, holding on to one well functioning protocol is suggested as this allows uninterrupted workflow, especially now when a well functioning and relativly simple protocol has been formulated. Also, the assembly of the samples to the DNA libraries should be done with more consideration, especially when there is time and expense related limitations for conducting possible sequencing re-runs. For example, here all of the +3y juveniles and the 54 adults captured at Utsjoki with parentages assinged to the +0y juveniles were included in a single library, and this library was also the one prepared with the one multiplex protocol where the primers behaved unexpectedly resulting in overall lower genotyping success. Hence, data acquisition from highly important samples was somewhat impaired due to utilization of a protocol which applicability wasn't ensured rigorously enough that could have been prevented by coducting the first actual sequencing on a library containing samples mainly from larger sample groups, such as the +0y and +1y juveniles. In order to avoid similar unnecessry risk that may lead to gaps in the final dataset or require expensive re-runs, the samples from different groups should be distributed evenly in all libraries. In this way possible batch dependent failures in the workflow won't compromise the dataset and prevent adressing the study questions in any major way.

ACKNOWLEDGEMENTS

Special thanks for my supervisor Academy professor Craig Primmer who gave me this opportunity to fulfill my thesis project in his group on this highly intriguing topic, his valuable insight and guidance and his endless patience and support even when this project has been delayed and stayed put momentarily. Similarly I'd like to give my special thanks to my second supervisor Dr. Tutku Aykanat for his valuable help and guidance, especially on the field of data analytics and R programming language. Thanks also to the other members of the Evolution, Conservation, and Genomics group in Helsinki University who were eager to help and advise me on more minor issues and made my time in Helsinki more pleasant.

Thanks to Meri Lindqvist and Katja Salminen from the Center of Evolutionary Applications in University of Turku who helped and advised me during my lab work.

Thanks also to my family and to all of my dear friends who have helped and supported me during this project and especially to the ones in Helsinki who also shared their homes with me during my time there.

REFERENCES

Aas, Øystein, Anders Klemetsen, Sigurd Einum, and Jostein Skurdal. 2011. *Atlantic Salmon Ecology (1st Edition).* https://books.google.com/books?id=9lMZnUdUGZUC&pg=PA240.

Albert, Arianne Y.K., Sterling Sawaya, Timothy H. Vines, Anne K. Knecht, Craig T. Miller, Brian R. Summers, Sarita Balabhadra, David M. Kingsley, and Dolph Schluter. 2008. "The Genetics of Adaptive Shape Shift in Stickleback: Pleiotropy and Effect Size." *Evolution* 62 (1): 76–85. https://doi.org/10.1111/j.1558-5646.2007.00259.x.

Aubin-Horth, Nadia, and Julian J. Dodson. 2004. "Influence of Individual Body Size and Variable Thresholds on the Incidence of a Sneaker Male Reproductive Tactic in Atlantic Salmon." *Evolution* 58 (1): 136–44. https://doi.org/10.1111/j.0014-3820.2004.tb01580.x.

Aykanat, T., M. Lindqvist, V. L. Pritchard, and C. R. Primmer. 2016. "From Population Genomics to Conservation and Management: A Workflow for Targeted Analysis of Markers Identified Using Genome-Wide Approaches in Atlantic Salmon Salmo Salar." *Journal of Fish Biology* 89 (6): 2658–79. https://doi.org/10.1111/jfb.13149.

Barrett, Rowan D.H., and Hopi E Hoekstra. 2011. "Molecular Spandrels: Tests of Adaptation at the Genetic Level." *Nature Reviews Genetics*. https://doi.org/10.1038/nrg3015.

Barson, Nicola J, Tutku Aykanat, Kjetil Hindar, Matthew Baranski, Geir H Bolstad, Peder Fiske, Céleste Jacq, et al. 2015. "Sex-Dependent Dominance at a Single Locus Maintains Variation in Age at Maturity in Salmon." *Nature* 528. https://doi.org/10.1038/nature16062.

Bonduriansky, Russell, and Stephen F. Chenoweth. 2009. "Intralocus Sexual Conflict." *Trends in Ecology & Evolution* 24 (5): 280–88. https://doi.org/10.1016/j.tree.2008.12.005.

Brownie, Jannine, Susan Shawcross, Jane Theaker, David Whitcombe, Richard Ferrie, Clive Newton, and Stephen Little. 1997. "The Elimination of Primer-Dimer Accumulation in PCR." *Nucleic Acids Research* 25 (16): 3235–41. https://doi.org/10.1093/nar/25.16.3235.

Chapman, G., J. Bangham, L. Rowe, and T. Arnqvist. 2003. "Sexual Conflict." *Trends in Ecology and Evolution* 18 (1): 41–48. https://doi.org/10.1016/S0169-5347(02)00004-6.

Charlesworth, Brian, and Deborah Charlesworth. 1978. "Model for Evolution of Dioecy and Gynodioecy." *The American Naturalist* 112 (988): 975–97.

Clarke, Geraldine M, Carl A. Anderson, Fredrik H. Pettersson, Lon R. Cardon, Andrew P. Morris, and Krina T. Zondervan. 2011. "Basic Statistical Analysis in Genetic Case-Control Studies." *Nature Protocols* 6 (February): 121. https://doi.org/10.1038/nprot.2010.182.

Coates, Brad S., Douglas V. Sumerford, Nicholas J. Miller, Kyung S. Kim, Thomas W. Sappington, Blair D. Siegfried, and Leslie C. Lewis. 2009. "Comparative Performance of Single Nucleotide Polymorphism and Microsatellite Markers for Population Genetic Analysis." *Journal of Heredity* 100 (5): 556–64. https://doi.org/10.1093/jhered/esp028.

Connallon, Tim, and Andrew G. Clark. 2014. "Balancing Selection in Species with Separate Sexes: Insights from Fisher's Geometric Model." *Genetics* 197 (3): 991–1006. https://doi.org/10.1534/genetics.114.165605.

Conte, I., J. M. Ruiz, N. Tabanera, R. Marco-Ferreres, P. Bovolenta, L. Beccari, and E. Cisneros. 2010. “Proper Differentiation of Photoreceptors and Amacrine Cells Depends on a Regulatory Loop between NeuroD and Six6.” *Development* 137 (14): 2307–17. https://doi.org/10.1242/dev.045294.

Cousminer, Diana L., Diane J. Berry, Nicholas J. Timpson, Wei Ang, Elisabeth Thiering, Enda M. Byrne, H. Rob Taal, et al. 2013. “Genome-Wide Association and Longitudinal Analyses Reveal Genetic Loci Linking Pubertal Height Growth, Pubertal Timing and Childhood Adiposity.” *Human Molecular Genetics* 22 (13): 2735–47. https://doi.org/10.1093/hmg/ddt104.

Czorlich, Yann, Tutku Aykanat, Jaakko Erkinaro, Panu Orell, and Craig R Primmer. 2018. “Rapid Sex-Specific Evolution of Age at Maturity Is Shaped by Genetic Architecture in Atlantic Salmon.” *BioRxiv*, 317255. https://doi.org/10.1101/317255.

Devlin, B, and Kathryn Roeder. 2004. “Genomic Control for Association Studies.” *International Biometric Society* 55 (4): 997–1004. http://www.jstor.org/stable/2533712.

Ellegren, Hans, and John Parsch. 2007. “The Evolution of Sex-Biased Genes and Sex-Biased Gene Expression.” *Nature Reviews Genetics* 8 (9): 689–98. https://doi.org/10.1038/nrg2167.

Ellmén, Mikko. 2015. “Atlantinlohen (Salmo Salar) Lisääntymismenestys Tenojoen Osapopulaatiossa: Onko Paikallisilla Kaloilla Kotikenttäetu?”.

Emlen, Stephen T, and Lewis W Oring. 1977. “Ecology, Sexual Selection, and the Evolution of Mating System.” *Science*. https://doi.org/10.1126/science.327542.

Erkinaro, Jaakko; Czorlich, Yann; Orell, Panu; Kuusela, Jorma; Falkegård, Morten; Länsman, Maija; Pulkkinen, Henni; Primmer, Craig R.; Niemelä, Eero. 2018. “Life History Variation across Four Decades in a Diverse Population Complex of Atlantic Salmon in a Large Subarctic River.” *Canadian Journal of Fisheries and Aquatic Sciences* 76 (1): 42–55. https://doi.org/10.1139/cjfas-2017-0343.

Erkinaro, Jaakko, Yann Czorlich, Panu Orell, Jorma Kuusela, Morten Falkegård, Maija Länsman, Henni Pulkkinen, Craig R Primmer, and Eero Niemelä. 2018. “Life History Variation across Four Decades in a Diverse Population Complex of Atlantic Salmon in a Large Subarctic River.” *Canadian Journal of Fisheries and Aquatic Sciences*, cjfas-2017-0343. https://doi.org/10.1139/cjfas-2017-0343.

Evans, Melissa L., Mélanie Dionne, Kristina M. Miller, and Louis Bernatchez. 2012. “Mate Choice for Major Histocompatibility Complex Genetic Divergence as a Bet-Hedging Strategy in the Atlantic Salmon (Salmo Salar).” *Proceedings of the Royal Society B: Biological Sciences* 279 (1727): 379–86. https://doi.org/10.1098/rspb.2011.0909.

Fan, WF, YC Liu, S Parimoo, and SM Weissman. 1995. “Olfactory Receptor-like Genes Are Located in Human MHC.” *Genomics* 27 (1): 119–23. https://doi.org/https://doi.org/10.1006/geno.1995.1013.

Fleming, Ian A. 1996. “Reproductive Strategies of Atlantic Salmon : Ecology and Evolution.” *Reviews in Fish Biology and Fisheries* 16 (6): 379–416.

Fleming, Ian A. 1998. “Pattern and Variability in the Breeding System of Atlantic Salmon (Salmo Salar), with Comparisons to Other Salmonids.” *Canadian Journal of Fisheries and Aquatic Sciences* 55 (S1): 59–76. https://doi.org/10.1139/d98-009.

Fry, James D. 2010. “The Genomic Location of Sexually Antagonistic Variation: Some Cautionary Comments.” *Evolution* 64 (5): 1510–16. https://doi.org/10.1111/j.1558-5646.2009.00898.x.

Garcia De Leaniz, C., I. A. Fleming, S. Einum, E. Verspoor, W. C. Jordan, S.

Consuegra, N. Aubin-Horth, et al. 2007. “A Critical Review of Adaptive Genetic Variation in Atlantic Salmon: Implications for Conservation.” *Biological Reviews* 82 (2): 173–211. https://doi.org/10.1111/j.1469-185X.2006.00004.x.

Gibson, J. R., A. K. Chippindale, and W. R. Rice. 2002. “The X Chromosome Is a Hot Spot for Sexually Antagonistic Fitness Variation.” *Proceedings of the Royal Society B: Biological Sciences* 269 (1490): 499–505. https://doi.org/10.1098/rspb.2001.1863.

Goodwin, Sara, John D McPherson, and W. Richard McCombie. 2016. “Coming of Age: Ten Years of next-Generation Sequencing Technologies.” *Nature Reviews Genetics*. https://doi.org/10.1038/nrg.2016.49.

Gutierrez, Alejandro P., Krzysztof P. Lubieniecki, Steve Fukui, Ruth E. Withler, Bruce Swift, and William S. Davidson. 2014. “Detection of Quantitative Trait Loci (QTL) Related to Grilsing and Late Sexual Maturation in Atlantic Salmon (Salmo Salar).” *Marine Biotechnology* 16 (1): 103–10. https://doi.org/10.1007/s10126-013-9530-3.

Halperin, Daniel S., Calvin Pan, Aldons J. Lusis, and Peter Tontonoz. 2013. “Vestigial-like 3 Is an Inhibitor of Adipocyte Differentiation.” *Journal of Lipid Research* 54 (2): 473–81. https://doi.org/10.1080/00036846.2014.937040.

Hansen, LP., N. Jonsson, and B. Jonsson. 1993. “Oceanic Migration in Homing Atlantic Salmon,” 927–41. https://doi.org/10.1006/anbe.1993.1112.

Hansen, LP., and TR. Quinn. 1998. “The Marine Phase of the Atlantic Salmon (Salmo Salar) Life Cycle, with Comparisons to Pacific Salmon.” *Canadian Journal Of Fisheries And Aquatic Sciences* 55 (1): 104–18. https://doi.org/10.1139/d98-010.

Heggenes, Jan. 1990. “Habitat Utilization and Preferences in Juvenile Atlantic Salmon (Salmo Salar) in Streams.” *Regulated Rivers: Research & Management* 5 (4): 341–54. https://doi.org/10.1002/rrr.3450050406.

Heinimaa, S., and J. Erkinaro. 2004. “Characteristics of Mature Male Parr in the Northernmost Atlantic Salmon Populations.” *Journal of Fish Biology* 64 (1): 219–26. https://doi.org/10.1046/j.1095-8649.2004.00308.x.

Heinimaa, Sirkka and Petri Heinimaa. 2004. “Effect of the Female Size on Egg Quality and Fecundity of the Wild Atlantic Salmon in the Sub-Arctic River Teno.” *Boreal Environment Research* 9 (1): 55–62. https://doi.org/10.1043/2009-0503-RA.1 [pii]\r10.1043/2009-0503-RA.1.

Howe, Glenn T., Sally N. Aitken, David B. Neale, Kathleen D. Jermstad, Nicholas C. Wheeler, and Tony HH. Chen. 2003. “From Genotype to Phenotype: Unraveling the Complexities of Cold Adaptation in Forest Trees.” *Canadian Journal of Botany* 81 (12): 1247–66. https://doi.org/10.1139/b03-141.

Hunt, John, Casper J. Breuker, Jennifer A. Sadowski, and Allen J. Moore. 2009. “Male-Male Competition, Female Mate Choice and Their Interaction: Determining Total Sexual Selection.” *Journal of Evolutionary Biology* 22 (1): 13–26. https://doi.org/10.1111/j.1420-9101.2008.01633.x.

Hutchings, Jeffrey A. and Ransom A. Myers. 1987. “Escalation of an Asymmetric Contest: Mortality Resulting from Mate Competition in Atlantic Salmon, *Salmo Salar*.” *Canadian Journal of Zoology* 65 (February 2011): 766–68. https://doi.org/10.1139/z87-120.

Järvi, Torbjörn. 1990. “The Effects of Male Dominance, Secondary Sexual Characteristics and Female Mate Choice on the Mating Success of Male Atlantic Salmon Salmo Salar.” *Ethology* 84 (2): 123–32. https://doi.org/10.1111/j.1439-0310.1990.tb00789.x.

Jean, Dominique, Gilbert Bernier, and Peter Gruss. 1999. “Six6 (Optx2) Is a Novel

Murine Six3-Related Homeobox Gene That Demarcates the Presumptive Pituitary/Hypothalamic Axis and the Ventral Optic Stalk." *Mechanisms of Development* 84 (1–2): 31–40. https://doi.org/10.1016/S0925-4773(99)00068-4.

Jennions, Michael D., and Marion Petrie. 2000. "Why Do Females Mate Multiply? A Review of the Genetic Benefits." *Biological Reviews* 75 (1): 21–64. https://doi.org/10.1111/j.1469-185X.1999.tb00040.x.

Johansen, Morten, Jaakko Erkinaro, Eero Niemelä, Tor G. Heggberget, Martin A. Svenning, and Sturla Brørs. 2016. "Status of the River Tana Salmon Populations 2016. Report of the Working Group on Salmon Monitoring and Research in the Tana River System."

Johnston, Susan E., Panu Orell, Victoria L. Pritchard, Matthew P. Kent, Sigbjørn Lien, Eero Niemelä, Jaakko Erkinaro, and Craig R. Primmer. 2014. "Genome-Wide SNP Analysis Reveals a Genetic Basis for Sea-Age Variation in a Wild Population of Atlantic Salmon (Salmo Salar)." *Molecular Ecology* 23 (14): 3452–68. https://doi.org/10.1111/mec.12832.

Jones, JW, and JH Orton. 1940. "The Paedogenetic Male Cycle in Salmo Salar L." *Proceedings of the Royal Society B: Biological Sciences* 128 (853): 485–99. https://doi.org/10.1098/rspb.1940.0022.

Jonsson, B., and N. Jonsson. 2009. "A Review of the Likely Effects of Climate Change on Anadromous Atlantic Salmon Salmo Salar and Brown Trout Salmo Trutta, with Particular Reference to Water Temperature and Flow." *Journal of Fish Biology* 75 (10): 2381–2447. https://doi.org/10.1111/j.1095-8649.2009.02380.x.

Jonsson, N., B. Jonsson, and L. P. Hansen. 1997. "Changes in Proximate Composition and Estimates of Energetic Costs During Upstream Migration and Spawning in Atlantic Salmon Salmo Salar." *The Journal of Animal Ecology* 66 (3): 425. https://doi.org/10.2307/5987.

Kazakov, R. V. 1981. "Peculiarities of Sperm Production by Anadromous and Parr Atlantic Salmon (Salmo Salar L.) and Fish Cultural Characteristics of Such Sperm." *Journal of Fish Biology* 18 (1): 1–8. https://doi.org/10.1111/j.1095-8649.1981.tb03753.x.

Keeley, E R, and JWA Grant. 2011. "Allometry of Diet Selectivity in Juvenile Atlantic Salmon (Salmo Salar)." *Canadian Journal of Fisheries and Aquatic Sciences* 54 (8): 1894–1902. https://doi.org/10.1139/f97-096.

Kekäläinen, Jukka, and Jonathan P. Evans. 2018. "Gamete-Mediated Mate Choice: Towards a More Inclusive View of Sexual Selection." *Proceedings of the Royal Society B: Biological Sciences* 285 (1883). https://doi.org/10.1098/rspb.2018.0836.

Kirkpatrick, Mark. 1982. "Sexual Selection and The Evolution of Female Choice." *Evolution* 36 (1): 1–12. https://doi.org/10.2307/2407961.

Klemetsen, A, P. A. Amundsen, J. B. Dempson, B Jonsson, N. Jonsson, M. F. O'Connell, and E. Mortensen. 2003. "Atlantic Salmon Salmo Salar L., Brown Trout Salmo Trutta L. and Arctic Charr Salvelinus Alpinus (L.): A Review of Aspects of Their Life Histories." *Ecology of Freshwater Fish* 12 (1): 1–59. https://doi.org/10.1034/j.1600-0633.2003.00010.x.

Lande, Russell. 1980. "Sexual Dimorphism, Sexual Selection, and Adaptation in Polygenic Characters." *Evolution* 34 (2): 292. https://doi.org/10.2307/2407393.

Lander, ES. S. and NJ. J. Schork. 1996. "Genetic Dissection of Complex Traits." *Nature Genetics* 12 (4): 355-356; author reply 357-358. https://doi.org/10.1016/S0065-2660(07)00419-1.

Landry, C., D. Garant, P. Duchesne, and L. Bernatchez. 2001. "'Good Genes as Heterozygosity': The Major Histocompatibility Complex and Mate Choice in

Atlantic Salmon (Salmo Salar)." *Proceedings of the Royal Society B: Biological Sciences* 268 (1473): 1279–85. https://doi.org/10.1098/rspb.2001.1659.
Larder, R., D. D. Clark, N. L. G. Miller, and P. L. Mellon. 2011. "Hypothalamic Dysregulation and Infertility in Mice Lacking the Homeodomain Protein Six6." *Journal of Neuroscience* 31 (2): 426–38. https://doi.org/10.1523/jneurosci.1688-10.2011.
Lepais, Olivier, Aurélie Manicki, Stéphane Glise, Mathieu Buoro, and Agnès Bardonnet. 2017. "Genetic Architecture of Threshold Reaction Norms for Male Alternative Reproductive Tactics in Atlantic Salmon (Salmo Salar L.)." *Scientific Reports* 7 (September 2016): 1–13. https://doi.org/10.1038/srep43552.
Lerceteau, Estelle, Christophe Plomion, and Bengt Andersson. 2000. "AFLP Mapping and Detection of Quantitative Trait Loci (QTLs) for Economically Important Traits in Pinus Sylvestris: A Preliminary Study." *Molecular Breeding* 6 (5): 451–58. https://doi.org/10.1023/A:1026548716320.
MacKay, Trudy F.C., Eric A. Stone, and Julien F. Ayroles. 2009. "The Genetics of Quantitative Traits: Challenges and Prospects." *Nature Reviews Genetics* 10 (8): 565–77. https://doi.org/10.1038/nrg2612.
Mackay, Trudy F C. 2001. "The Genetic Architecture of Quantitative Traits." *Annual Review of Genetics* 35: 303–39. https://doi.org/10.1146/annurev.genet.35.102401.090633.
Mayden, Richard L, Kevin L Tang, Kevin W Conway, Joerg Freyhof, Sarah Chamberlain, Miranda Haskins, Leah Schneider, et al. 2007. "Phylogenetic Relationships of Danio within the Order Cypriniformes : A Framework for Comparative and Evolutionary Studies of a Model Species." *Journal of Experimental Zoology. Part B, Molecular and Developmental Evolution* 308B (5): 642–54. https://doi.org/10.1002/jez.b.21175.
McCormick, Stephen D., Lars P. Hansen, Thomas P. Quinn, and Richard L. Saunders. 1998. "Movement, Migration, and Smolting of Atlanticsalmon (Salmo Salar)." *Canadian Journal of Fisheries and Aquatic Sciences* 55: 77–92. https://doi.org/10.1007/978-3-319-25582-8_180007.
Merilä, Juha. 2013. "Nine-Spined Stickleback (Pungitius Pungitius): An Emerging Model for Evolutionary Biology Research." *Annals of the New York Academy of Sciences* 1289 (1): 18–35. https://doi.org/10.1111/nyas.12089.
Meyer, Axel, Romulus Abila, Dickson Otieno Owiti, Millicent Florence Ndonga, Marta Barluenga, and Walter Salzburger. 2010. "The Role of the Yala Swamp Lakes in the Conservation of Lake Victoria Region Haplochromine Cichlids: Evidence from Genetic and Trophic Ecology Studies." *Lakes & Reservoirs: Research & Management* 13 (2): 95–104. https://doi.org/10.1111/j.1440-1770.2008.00366.x.
Mjølnerød, I.B., I.A. Fleming, U.H. Refseth, and K. Hindar. 1998. "Mate and Sperm Competition during Multiple-Male Spawnings of Atlantic Salmon." *Canadian Journal of Zoology* 76 (1): 70–75. https://doi.org/10.1139/cjz-76-1-70.
Mobley, Kenyon B., Hanna Granroth-Wilding, Mikko Ellmen, Juha-Pekka Vähä, Tutku Aykanat, Susan E. Johnston, Panu Orell, Jaakko Erkinaro, and Craig R. Primmer. 2019. "Home Ground Advantage: Local Atlantic Salmon Have Higher Reproductive Fitness than Dispersers in the Wild." *Science Advances* 5 (2): eaav1112. https://doi.org/10.1126/sciadv.aav1112.
Naish, KA, and JJ Hard. 2008. "Bridging the Gap between the Genotype and the Phenotype: Linking Genetic Variation, Selection and Adaptation in Fishes." *Fish and Fisheries* 9 (4): 396–422. https://doi.org/10. 1111/j. 1467-2979. 2008. 00302. x.

Neff, Bryan D., and Trevor E. Pitcher. 2005. “Genetic Quality and Sexual Selection: An Integrated Framework for Good Genes and Compatible Genes.” *Molecular Ecology* 14 (1): 19–38. https://doi.org/10.1111/j.1365-294X.2004.02395.x.

Olse, K A N. 1998. “MHC and Kin Discrimination in Juvenile Arctic Charr, Salvelinus Alpinus (L.).” *Animal Behaviour*, 319–27.

Parker, G. A. 1978. “Selection on Non-Random Fusion of Gametes during the Evolution of Anisogamy.” *Journal of Theoretical Biology* 73 (1): 1–28. https://doi.org/10.1016/0022-5193(78)90177-7.

Parrish, Donna L, Robert J Behnke, Stephen R Gephard, Stephen D McCormick, and Gordon H Reeves. 2011. “Why Aren’t There More Atlantic Salmon (Salmo Salar)?” *Canadian Journal of Fisheries and Aquatic Sciences* 55: 281–87. http://www.nrcresearchpress.com/doi/abs/10.1139/d98-012#.UaNZ70D0GSo.

Penn, D. J., K. Damjanovich, and W. K. Potts. 2002. “MHC Heterozygosity Confers a Selective Advantage against Multiple-Strain Infections.” *Proceedings of the National Academy of Sciences* 99 (17): 11260–64. https://doi.org/10.1073/pnas.162006499.

Piertney, S. B., and M. K. Oliver. 2006. “The Evolutionary Ecology of the Major Histocompatibility Complex.” *Heredity* 96 (1): 7–21. https://doi.org/10.1038/sj.hdy.6800724.

Pritchard, Victoria L., Hannu Mäkinen, Juha Pekka Vähä, Jaakko Erkinaro, Panu Orell, and Craig R. Primmer. 2018. “Genomic Signatures of Fine-Scale Local Selection in Atlantic Salmon Suggest Involvement of Sexual Maturation, Energy Homeostasis and Immune Defence-Related Genes.” *Molecular Ecology* 27 (11): 2560–75. https://doi.org/10.1111/mec.14705.

Reinton, Nils, Philippe Collas, Trine B. Haugen, Bjørn S. Skålhegg, Vidar Hansson, Tore Jahnsen, and Kjetil Taskén. 2000. “Localization of a Novel Human A-Kinase-Anchoring Protein, HAKAP220, during Spermatogenesis.” *Developmental Biology* 223 (1): 194–204. https://doi.org/10.1006/dbio.2000.9725.

Rice, William. 1984. “Sex Chromosomes and the Evolution of Sexual Dimorphism.” *Society for the Study of Evolution* 38 (4): 735–42. https://doi.org/10.2307/2408385.

Rice, William R, and Brett Holland. 1997. “The Enemies within : Intergenomic Conflict, Interlocus Contest Evolution (ICE), and the Intraspecific Red Queen.” *Behavioral Ecology and Sociobiology* 41 (1): 1–10.

Rothberg, Jonathan M, Wolfgang Hinz, Todd M Rearick, Jonathan Schultz, William Mileski, Mel Davey, John H Leamon, et al. 2011. “An Integrated Semiconductor Device Enabling Non-Optical Genome Sequencing.” *Nature* 475 (July): 348. https://doi.org/10.1038/nature10242.

Saunders, Richard L., Eugene B. Henderson, and Brian D. Glebe. 1982. “Precocious Sexual Maturation and Smoltification in Atlantic Salmon (Salmo Salar).” *Aquaculture* 28: 211–29. https://doi.org/10.1016/0044-8486(82)90024-2.

Saunders, Richard L, and Charles B Schom. 2008. “Importance of the Variation in Life History Parameters of Atlantic Salmon (Salmo Salar).” *Canadian Journal of Fisheries and Aquatic Sciences* 42 (3): 615–18. https://doi.org/10.1139/f85-080.

Shine, Richard. 1989. “Ecological Causes for the Evolution of Sexual Dimorphism: A Review of the Evidence.” *The Quarterly Review of Biology* 64 (4): 419–61. https://doi.org/10.1086/416458.

Simpson, A. L. 1992. “Differences in Body Size and Lipid Reserves between Maturing and Nonmaturing Atlantic Salmon Parr, Salmo Salar L.” *Canadian Journal of Zoology* 70 (9): 1737–42. https://doi.org/10.1139/z92-241.

Taranger, Geir Lasse, Manuel Carrillo, Rüdiger W. Schulz, Pascal Fontaine, Silvia

Zanuy, Alicia Felip, Finn Arne Weltzien, et al. 2010. “Control of Puberty in Farmed Fish.” *General and Comparative Endocrinology* 165 (3): 483–515. https://doi.org/10.1016/j.ygcen.2009.05.004.

Tentelier, Cédric, Olivier Lepais, Nicolas Larranaga, Aurélie Manicki, Frédéric Lange, and Jacques Rives. 2016. “Sexual Selection Leads to a Tenfold Difference in Reproductive Success of Alternative Reproductive Tactics in Male Atlantic Salmon.” *Science of Nature* 103 (5). https://doi.org/10.1007/s00114-016-1372-1.

Thibert-Plante, Xavier, and Sergey Gavrilets. 2013. “Evolution of Mate Choice and the So-Called Magic Traits in Ecological Speciation.” *Ecology Letters* 16 (8): 1004–13. https://doi.org/10.1111/ele.12131.

Thomaz, D., E. Beall, and T. Burke. 1997. “Alternative Reproductive Tactics in Atlantic Salmon: Factors Affecting Mature Parr Success.” *Proceedings of the Royal Society B: Biological Sciences* 264 (1379): 219–26. https://doi.org/10.1098/rspb.1997.0031.

Thorpe, JE, MS Miles, and DS Keay. 1984. “Developmental Rate, Fecundity And Egg Size in Atlanticsalmon, Salmo Salar L.” *Aquaculture* 43 (1–3): 289–305. https://doi.org/10.1016/0044-8486(84)90030-9.

Tregenza, T, and N Wedell. 2000. “Genetic Compatibility , Mate Choice and Patterns of Parentage : Invited Review.” *Molecular Ecology* 9 (8): 1013–27. https://doi.org/10.1046/j.1365-294x.2000.00964.x.

Trombley, Susanne, Arshi Mustafa, and Monika Schmitz. 2014. “Regulation of the Seasonal Leptin and Leptin Receptor Expression Profile during Early Sexual Maturation and Feed Restriction in Male Atlantic Salmon, Salmo Salar L., Parr.” *General and Comparative Endocrinology* 204: 60–70. https://doi.org/10.1016/j.ygcen.2014.04.033.

Vladic, TV, and Torbjörn Järvi. 2001. “Sperm Quality in the Alternative Reproductive Tactics of Atlantic Salmon: The Importance of the Loaded Raffle Mechanism.” *Proceedings of the Royal Society B: Biological Sciences* 268 (1483): 2375–81. https://doi.org/10.1098/rspb.2001.1768.

Ward, RD. 2000. “Genetics in Fisheries Management.” *Hydrobiologia* 420 (1): 191–201. http://dx.doi.org/10.1023/A:1003928327503.

Watanabe, Nanako, Kae Itoh, Makoto Mogi, Yuichiro Fujinami, Daisuke Shimizu, Hiroshi Hashimoto, Susumu Uji, Hayato Yokoi, and Tohru Suzuki. 2012. “Circadian Pacemaker in the Suprachiasmatic Nuclei of Teleost Fish Revealed by Rhythmic Period2 Expression.” *General and Comparative Endocrinology* 178 (2): 400–407. https://doi.org/10.1016/j.ygcen.2012.06.012.

Yang, Jian, Michael N. Weedon, Shaun Purcell, Guillaume Lettre, Karol Estrada, Cristen J. Willer, Albert V. Smith, et al. 2011. “Genomic Inflation Factors under Polygenic Inheritance.” *European Journal of Human Genetics* 19 (7): 807–12. https://doi.org/10.1038/ejhg.2011.39.

Yeates, Sarah E., Sigurd Einum, Ian A. Fleming, Hendrik Jan Megens, René J.M. Stet, Kjetil Hindar, William V. Holt, Katrien J.W. Van Look, and Matthew J.G. Gage. 2009. “Atlantic Salmon Eggs Favour Sperm in Competition That Have Similar Major Histocompatibility Alleles.” *Proceedings of the Royal Society B: Biological Sciences* 276 (1656): 559–66. https://doi.org/10.1098/rspb.2008.1257.

APPENDIX A: Detailed final PCR-1 & -2 reactions and reaction components

Table 1. Detailed final PCR-1 reaction components and conditions.

Reaction component	Multiplex 1 volume (µl)	Multiplex 2 volume (µl)	Temperature (°C)	Time	Cycles
QMP 2x*	5.5	5.5	95	15 min	
Primer mix**	1.2	2.7	95	30 s	
H2O	2.8	1.3	58	1 min	x 7
DNA	1.5	1.5	72	45 s	
Total	11	11	95	30 s	
			62	1 min	x 14
			72	45 s	

* 2x QIAGEN Multiplex PCR Master Mix
** Different set of primers were included in the primer mixes of different multiplexes

Table 2. Detailed final PCR-2 reaction components and conditions. The sequences of the customised reverse primers with highlighted barcodes are attached below**.** For SequalPrep™ Normalization protocol the volumes of the reaction components were adjusted so that the total reaction volume was 20 µl and three extra cycles were added.

Reaction component	volume (µl)	Temperature (°C)	Time	Cycles
QMP 2x	7.5	95	15 min	
Ion-A_IonX_xxx_Uni-T7*	4.2	98	20 s	
Ion-trP1_xxx_Uni-tagR**	1.05	60	30 s	x 15
H2O	0.25	72	30 s	
DNA (purified PCR-1 product***)	2	72	5 min	
Total	15	10	1 min	

* Ion Xpress™ Barcode forward primers with 1-96 unique barcode sequences as well as Ion A adapter - sequence allowing the binding of amplicons to the ISPs.
** Customized reverse barcodes with 1-8 unique barcode sequences (sequences listed below)
*** PCR-1 amplicon DNA after purification with SPRI-beads

Ion-trP1_i01_Uni-tagR	CCTCTCTATGGGCAGTCGGTGAT**CGTGAT**CATTAAGTTCCCATTA
Ion-trP1_i02_Uni-tagR	CCTCTCTATGGGCAGTCGGTGAT**ACATCG**CATTAAGTTCCCATTA
Ion-trP1_i03_Uni-tagR	CCTCTCTATGGGCAGTCGGTGAT**GCCTAA**CATTAAGTTCCCATTA
Ion-trP1_i04_Uni-tagR	CCTCTCTATGGGCAGTCGGTGAT**TGGTCA**CATTAAGTTCCCATTA
Ion-trP1_i05_Uni-tagR	CCTCTCTATGGGCAGTCGGTGAT**CACTGT**CATTAAGTTCCCATTA
Ion-trP1_i06_Uni-tagR	CCTCTCTATGGGCAGTCGGTGAT**ATTGGC**CATTAAGTTCCCATTA
Ion-trP1_i07_Uni-tagR	CCTCTCTATGGGCAGTCGGTGAT**GATCTG**CATTAAGTTCCCATTA
Ion-trP1_i08_Uni-tagR	CCTCTCTATGGGCAGTCGGTGAT**TCAAGT**CATTAAGTTCCCATTA

APPENDIX B: Known maiting pairs and the offspring born for these pairs

Table 1. The known mating pairs of spawning Atlantic salmon of lower Utsjoki region population of 2011 and the amount of offspring born for each mating pair. The *vgll3*$_{TOP}$ genotype of adults are indicated in parenthesis after the name of the individual.

			Females					
			Adult_2011_22 (*LE*)	Adult_2011_29 (*LE*)	Adult_2011_37 (*LE*)	Adult_2011_39 (*LL*)	Adult_2011_52 (*LL*)	Adult_2011_53 (*LL*)
Males	Adult_2011_2	(*LL*)	2	-	-	-	-	-
	Adult_2011_6	(*LE*)	-	-	-	-	2	-
	Adult_2011_8	(*EE*)	-	-	-	1	-	-
	Adult_2011_9	(*LE*)	-	-	-	-	-	3
	Adult_2011_15	(*EE*)	-	-	-	-	-	5
	Adult_2011_17	(*LL*)	-	1	1	-	-	-
	Adult_2011_18	(*EE*)	-	-	-	5	-	-
	Adult_2011_26	(*LE*)	9	-	-	-	-	-
	Adult_2011_36	(*LE*)	-	5	2	1	-	-
	Adult_2011_40	(*LL*)	-	-	-	5	-	-
	Adult_2011_42	(*LE*)	-	-	-	-	-	1

APPENDIX C: Segregation distortion in *vgll3* alleles in offspring born for known maiting pairs

Table 1. Expected Mendelian and observed offspring ratios born for known maiting pairs where both parents are heterozygous *vgll3* genotype. There was no significant deviation between the expected and observed offspring ratios (Fisher's exact test, P-value = 0.44)

	Expected	Observed
LL	3	5
LE	7	3
EE	4	6

Table 2. Expected Mendelian and observed offspring ratios born for known maiting pairs where one parent is *vgll3 LL* homozygous and the other heterozygous genotype. There was no significant deviation between the expected and observed offspring ratios (Fisher's exact test, P-value = 0.14)

	Expected	Observed
LL	5	1
LE	5	9

APPENDIX D: Change in the *vgll3* genotype frequencies in lower Utsjoki region juvenile cohort hatched on 2012

Table 1. P-values yielded by the Fisher's exact test when the *vgll3* genotype frequencies were compared between all individuals and males and females separately of different cohort years and within one cohort year between the sexes. No significant deviations was observed in any setting (P-value > 0.05).

		+0y 2012	+1y 2013	+2y 2014	+3y 2015
+0y 2012	All	-	0.43	0.29	0.66
	Female to female	-	0.48	0.72	0.21
	Male to male	-	0.77	0.16	0.81
Female to male (Within a cohort year sampling)		0.47	0.96	0.51	0.41

APPENDIX E: The sea-age phenotype, *vgll3* genotype and offspring count of the lower Utsjoki region adult Atlantic salmon from 2011.

Table 1. The sea-age phenotype, vgll3 genotype and offspring count of the adult lower Utsjoki region adult Atlantic salmon from 2011. The sea-age phenotype, *vgll3* genotype and offspring count of the lower Utsjoki region adult Atlantic salmon from 2011. Both the 54 adults captured from the spawning site and the 46 additional adults (add included in the id) assigned to the population via molecular markers are included. Even though the pedigree information (Appendix F) allows direct observation of the fitness of adult individuals of certain *vgll3* genotype, the amount of adults with information of known offspring available was notably small, and thus no conclusions of the fitness effects of certain genotypes could be drawn. Hence this approach was excluded from the actual analysis of this study.

ID	Sea-age	*vgll3* genotype	Offspring count	Sex
Adult_add_20111284	SW3	*LL*	NA	Female
Adult_add_20111412	2S1	*LL*	NA	Female
Adult_add_20111732	SW3	*LL*	NA	Female
Adult_add_20113033	SW3	*LL*	NA	Female
Adult_2011_39	SW4	*LL*	23	Female
Adult_2011_52	SW3	*LL*	6	Female
Adult_2011_53	SW3	*LL*	27	Female
Adult_add_20112600	SW3	*LE*	NA	Female
Adult_add_20112799	2S1	*LE*	NA	Female
Adult_add_2011709	SW3	*LE*	NA	Female
Adult_add_2011892	SW3	*LE*	NA	Female
Adult_2011_22	SW2	*LE*	20	Female
Adult_2011_28	SW3	*LE*	25	Female
Adult_2011_29	SW3	*LE*	23	Female
Adult_2011_37	SW4	*LE*	26	Female
Adult_add_20112802	SW1	*EE*	NA	Female
Adult_add_2011703	SW4	*EE*	NA	Female
Adult_add_20111697	SW4	*LL*	NA	Male
Adult_add_20111727	SW1	*LL*	NA	Male
Adult_add_20111730	SW1	*LL*	NA	Male
Adult_add_2011696	SW1	*LL*	NA	Male
Adult_add_201192	SW1	*LL*	NA	Male
Adult_2011_13	SW1	*LL*	NA	Male
Adult_2011_17	SW3	*LL*	10	Male
Adult_2011_2	SW1	*LL*	6	Male
Adult_2011_40	SW5	*LL*	33	Male
Adult_2011_44	SW1	*LL*	NA	Male
Adult_2011_7	SW3	*LL*	4	Male
Adult_add_20111041	SW1	*LE*	NA	Male
Adult_add_20111285	SW1	*LE*	NA	Male
Adult_add_20111305	SW3	*LE*	NA	Male
Adult_add_20111365	SW1	*LE*	NA	Male
Adult_add_20111462	SW1	*LE*	NA	Male

Adult_add_20111699	SW1	*LE*	NA	Male
Adult_add_20111714	SW1	*LE*	NA	Male
Adult_add_20111813	SW1	*LE*	NA	Male
Adult_add_20111814	SW1	*LE*	NA	Male
Adult_add_20111819	SW1	*LE*	NA	Male
Adult_add_20112020	SW1	*LE*	NA	Male
Adult_add_20112616	1S1	*LE*	NA	Male
Adult_add_201128	SW1	*LE*	NA	Male
Adult_add_20113031	SW1	*LE*	NA	Male
Adult_add_20113032	SW3	*LE*	NA	Male
Adult_add_2011713	SW1	*LE*	NA	Male
Adult_add_2011715	SW1	*LE*	NA	Male
Adult_add_2011869	SW1	*LE*	NA	Male
Adult_add_2011875	SW1	*LE*	NA	Male
Adult_add_201199	SW1	*LE*	NA	Male
Adult_2011_1	SW1	*LE*	NA	Male
Adult_2011_11	SW1	*LE*	1	Male
Adult_2011_12	SW1	*LE*	15	Male
Adult_2011_14	SW2	*LE*	2	Male
Adult_2011_16	SW2	*LE*	3	Male
Adult_2011_19	SW1	*LE*	NA	Male
Adult_2011_20	SW1	*LE*	2	Male
Adult_2011_21	SW1	*LE*	NA	Male
Adult_2011_23	SW1	*LE*	NA	Male
Adult_2011_24	SW1	*LE*	NA	Male
Adult_2011_25	SW1	*LE*	2	Male
Adult_2011_26	SW1	*LE*	16	Male
Adult_2011_27	SW1	*LE*	NA	Male
Adult_2011_30	SW1	*LE*	1	Male
Adult_2011_32	SW1	*LE*	NA	Male
Adult_2011_35	SW1	*LE*	NA	Male
Adult_2011_36	SW3	*LE*	33	Male
Adult_2011_38	SW1	*LE*	NA	Male
Adult_2011_41	SW1	*LE*	NA	Male
Adult_2011_42	SW1	*LE*	1	Male
Adult_2011_43	SW1	*LE*	5	Male
Adult_2011_45	SW1	*LE*	1	Male
Adult_2011_46	SW1	*LE*	NA	Male
Adult_2011_47	SW1	*LE*	NA	Male
Adult_2011_5	SW1	*LE*	NA	Male
Adult_2011_50	SW1	*LE*	NA	Male
Adult_2011_51	SW1	*LE*	2	Male
Adult_2011_54	SW1	*LE*	1	Male
Adult_2011_6	SW2	*LE*	7	Male
Adult_2011_9	SW1	*LE*	7	Male
Adult_add_201110	SW1	*EE*	NA	Male

Adult_add_2011100	SW1	*EE*	NA	Male
Adult_add_20111304	SW1	*EE*	NA	Male
Adult_add_20111409	1S1	*EE*	NA	Male
Adult_add_20111635	SW1	*EE*	NA	Male
Adult_add_20111642	SW2	*EE*	NA	Male
Adult_add_20111733	SW1	*EE*	NA	Male
Adult_add_20111804	1S1	*EE*	NA	Male
Adult_add_20111811	SW1	*EE*	NA	Male
Adult_add_20112798	SW2	*EE*	NA	Male
Adult_add_2011889	SW1	*EE*	NA	Male
Adult_2011_10	SW1	*EE*	NA	Male
Adult_2011_15	SW1	*EE*	6	Male
Adult_2011_18	SW1	*EE*	12	Male
Adult_2011_3	SW2	*EE*	NA	Male
Adult_2011_31	SW1	*EE*	NA	Male
Adult_2011_33	SW1	*EE*	1	Male
Adult_2011_34	SW1	*EE*	2	Male
Adult_2011_4	SW2	*EE*	NA	Male
Adult_2011_48	SW1	*EE*	1	Male
Adult_2011_49	SW1	*EE*	2	Male
Adult_2011_8	SW1	*EE*	1	Male

APPENDIX F: The pedigree information of the lower Utsjoki region adult Atlantic salmon of 2011 and the subsequent offspring hatched on 2012 (Ellmén 2015).

Table 1. The pedigree information of the lower Utsjoki region adult Atlantic salmon of 2011 and the subsequent offspring hatched on 2012. This pedigree was created by Ellmen (2015) based on the microsatellite markers.

Offspring ID	Dam ID	Sire ID
+0y_2012_4	Adult_2011_37	NA
+0y_2012_7	NA	Adult_2011_51
+0y_2012_9	NA	Adult_2011_40
+0y_2012_10	Adult_2011_29	NA
+0y_2012_14	Adult_2011_28	NA
+0y_2012_15	NA	Adult_2011_40
+0y_2012_19	NA	Adult_2011_40
+0y_2012_22	NA	Adult_2011_12
+0y_2012_25	Adult_2011_29	Adult_2011_36
+0y_2012_26	Adult_2011_29	NA
+0y_2012_27	Adult_2011_39	NA
+0y_2012_31	NA	Adult_2011_12
+0y_2012_65	NA	Adult_2011_2
+0y_2012_71	NA	Adult_2011_36
+0y_2012_80	NA	Adult_2011_40
+0y_2012_81	Adult_2011_39	Adult_2011_40
+0y_2012_85	Adult_2011_37	NA
+0y_2012_87	Adult_2011_37	NA
+0y_2012_90	Adult_2011_29	NA
+0y_2012_91	NA	Adult_2011_40
+0y_2012_92	NA	Adult_2011_25
+0y_2012_139	Adult_2011_29	NA
+0y_2012_141	NA	Adult_2011_2
+0y_2012_148	Adult_2011_28	NA
+0y_2012_149	Adult_2011_37	NA
+0y_2012_151	Adult_2011_39	Adult_2011_18
+0y_2012_154	Adult_2011_28	NA
+0y_2012_155	Adult_2011_28	NA
+0y_2012_156	NA	Adult_2011_36
+0y_2012_160	Adult_2011_28	NA
+0y_2012_164	Adult_2011_29	NA
+0y_2012_166	Adult_2011_28	NA
+0y_2012_174	Adult_2011_39	Adult_2011_8
+0y_2012_176	Adult_2011_39	NA
+0y_2012_40	Adult_2011_37	NA
+0y_2012_42	NA	Adult_2011_12
+0y_2012_44	Adult_2011_39	NA
+0y_2012_50	Adult_2011_37	Adult_2011_36

+0y_2012_51	Adult_2011_39	NA
+0y_2012_52	NA	Adult_2011_18
+0y_2012_56	NA	Adult_2011_43
+0y_2012_57	NA	Adult_2011_36
+0y_2012_58	Adult_2011_53	Adult_2011_9
+0y_2012_61	NA	Adult_2011_40
+0y_2012_62	Adult_2011_37	NA
+0y_2012_63	NA	Adult_2011_16
+0y_2012_103	NA	Adult_2011_12
+0y_2012_111	NA	Adult_2011_2
+0y_2012_127	Adult_2011_28	NA
+0y_2012_128	NA	Adult_2011_34
+0y_2012_134	Adult_2011_37	NA
+0y_2012_178	Adult_2011_28	NA
+0y_2012_180	NA	Adult_2011_12
+0y_2012_181	Adult_2011_28	NA
+0y_2012_182	Adult_2011_39	Adult_2011_40
+0y_2012_183	NA	Adult_2011_36
+0y_2012_184	NA	Adult_2011_36
+0y_2012_187	Adult_2011_28	NA
+0y_2012_188	Adult_2011_39	NA
+0y_2012_190	Adult_2011_39	Adult_2011_40
+0y_2012_206	NA	Adult_2011_26
+0y_2012_213	NA	Adult_2011_12
+0y_2012_214	Adult_2011_22	Adult_2011_26
+0y_2012_217	Adult_2011_28	NA
+0y_2012_218	NA	Adult_2011_20
+0y_2012_219	Adult_2011_28	NA
+0y_2012_223	NA	Adult_2011_40
+0y_2012_226	Adult_2011_39	Adult_2011_40
+0y_2012_229	NA	Adult_2011_17
+0y_2012_233	NA	Adult_2011_25
+0y_2012_240	NA	Adult_2011_51
+0y_2012_246	NA	Adult_2011_12
+0y_2012_249	Adult_2011_22	NA
+0y_2012_252	Adult_2011_28	NA
+0y_2012_255	Adult_2011_28	NA
+0y_2012_256	Adult_2011_28	NA
+0y_2012_257	NA	Adult_2011_12
+0y_2012_258	NA	Adult_2011_43
+0y_2012_259	Adult_2011_28	NA
+0y_2012_260	Adult_2011_28	NA
+0y_2012_261	NA	Adult_2011_17
+0y_2012_290	Adult_2011_37	Adult_2011_36
+0y_2012_291	NA	Adult_2011_36
+0y_2012_293	Adult_2011_28	NA

+0y_2012_294	Adult_2011_28	NA
+0y_2012_295	NA	Adult_2011_40
+0y_2012_298	Adult_2011_22	NA
+0y_2012_303	Adult_2011_22	Adult_2011_26
+0y_2012_304	NA	Adult_2011_17
+0y_2012_306	NA	Adult_2011_36
+0y_2012_308	Adult_2011_37	NA
+0y_2012_318	Adult_2011_37	NA
+0y_2012_320	Adult_2011_39	Adult_2011_36
+0y_2012_321	NA	Adult_2011_11
+0y_2012_325	NA	Adult_2011_40
+0y_2012_326	Adult_2011_29	NA
+0y_2012_327	Adult_2011_39	NA
+0y_2012_328	Adult_2011_29	NA
+0y_2012_330	NA	Adult_2011_40
+0y_2012_331	NA	Adult_2011_26
+0y_2012_332	NA	Adult_2011_12
+0y_2012_335	NA	Adult_2011_36
+0y_2012_336	Adult_2011_39	NA
+0y_2012_341	NA	Adult_2011_36
+0y_2012_343	NA	Adult_2011_12
+0y_2012_345	NA	Adult_2011_17
+0y_2012_350	NA	Adult_2011_40
+0y_2012_353	Adult_2011_22	Adult_2011_26
+0y_2012_354	Adult_2011_37	NA
+0y_2012_355	NA	Adult_2011_12
+0y_2012_356	NA	Adult_2011_26
+0y_2012_358	NA	Adult_2011_9
+0y_2012_359	Adult_2011_53	NA
+0y_2012_362	NA	Adult_2011_26
+0y_2012_366	NA	Adult_2011_40
+0y_2012_370	Adult_2011_29	NA
+0y_2012_371	Adult_2011_22	Adult_2011_26
+0y_2012_372	Adult_2011_39	NA
+0y_2012_375	NA	Adult_2011_33
+0y_2012_376	Adult_2011_37	NA
+0y_2012_377	Adult_2011_29	Adult_2011_36
+0y_2012_378	NA	Adult_2011_40
+0y_2012_379	Adult_2011_29	NA
+0y_2012_380	NA	Adult_2011_40
+0y_2012_384	Adult_2011_29	NA
+0y_2012_385	NA	Adult_2011_36
+0y_2012_386	Adult_2011_22	NA
+0y_2012_389	NA	Adult_2011_17
+0y_2012_390	Adult_2011_29	NA
+0y_2012_392	Adult_2011_37	NA

+0y_2012_393	NA	Adult_2011_30
+0y_2012_395	NA	Adult_2011_36
+0y_2012_399	Adult_2011_22	NA
+0y_2012_401	NA	Adult_2011_18
+0y_2012_404	Adult_2011_29	Adult_2011_36
+0y_2012_405	Adult_2011_39	Adult_2011_18
+0y_2012_408	Adult_2011_28	NA
+0y_2012_409	Adult_2011_22	Adult_2011_26
+0y_2012_412	Adult_2011_22	NA
+0y_2012_413	Adult_2011_22	Adult_2011_26
+0y_2012_417	NA	Adult_2011_36
+0y_2012_418	Adult_2011_22	Adult_2011_26
+0y_2012_420	NA	Adult_2011_40
+0y_2012_422	NA	Adult_2011_26
+0y_2012_426	NA	Adult_2011_17
+0y_2012_427	NA	Adult_2011_43
+0y_2012_430	NA	Adult_2011_40
+0y_2012_431	NA	Adult_2011_40
+0y_2012_440	Adult_2011_22	Adult_2011_2
+0y_2012_441	NA	Adult_2011_36
+0y_2012_446	NA	Adult_2011_17
+0y_2012_447	NA	Adult_2011_36
+0y_2012_448	NA	Adult_2011_36
+0y_2012_449	NA	Adult_2011_17
+0y_2012_452	Adult_2011_37	NA
+0y_2012_455	Adult_2011_29	NA
+0y_2012_456	NA	Adult_2011_16
+0y_2012_457	Adult_2011_29	Adult_2011_36
+0y_2012_463	Adult_2011_22	NA
+0y_2012_468	Adult_2011_37	NA
+0y_2012_469	Adult_2011_29	NA
+0y_2012_470	Adult_2011_29	Adult_2011_17
+0y_2012_471	Adult_2011_22	Adult_2011_2
+0y_2012_475	Adult_2011_39	NA
+0y_2012_477	NA	Adult_2011_40
+0y_2012_479	NA	Adult_2011_40
+0y_2012_480	Adult_2011_29	NA
+0y_2012_266	Adult_2011_29	NA
+0y_2012_269	Adult_2011_37	NA
+0y_2012_270	Adult_2011_39	Adult_2011_18
+0y_2012_272	NA	Adult_2011_36
+0y_2012_274	Adult_2011_28	NA
+0y_2012_279	NA	Adult_2011_40
+0y_2012_287	NA	Adult_2011_18
+0y_2012_288	NA	Adult_2011_20
+0y_2012_489	Adult_2011_22	NA

+0y_2012_491	NA	Adult_2011_43
+0y_2012_497	Adult_2011_53	Adult_2011_15
+0y_2012_499	Adult_2011_39	NA
+0y_2012_504	Adult_2011_29	NA
+0y_2012_505	NA	Adult_2011_43
+0y_2012_511	Adult_2011_22	NA
+0y_2012_515	NA	Adult_2011_12
+0y_2012_519	NA	Adult_2011_2
+0y_2012_525	Adult_2011_37	NA
+0y_2012_527	NA	Adult_2011_48
+0y_2012_531	NA	Adult_2011_26
+0y_2012_532	Adult_2011_37	NA
+0y_2012_533	NA	Adult_2011_49
+0y_2012_534	NA	Adult_2011_18
+0y_2012_535	Adult_2011_29	Adult_2011_36
+0y_2012_537	NA	Adult_2011_40
+0y_2012_539	NA	Adult_2011_16
+0y_2012_544	NA	Adult_2011_9
+0y_2012_545	NA	Adult_2011_36
+0y_2012_553	NA	Adult_2011_7
+0y_2012_559	NA	Adult_2011_6
+0y_2012_560	NA	Adult_2011_36
+0y_2012_569	NA	Adult_2011_6
+0y_2012_571	Adult_2011_52	Adult_2011_6
+0y_2012_592	NA	Adult_2011_14
+0y_2012_593	Adult_2011_52	NA
+0y_2012_595	Adult_2011_53	NA
+0y_2012_606	Adult_2011_52	NA
+0y_2012_607	NA	Adult_2011_34
+0y_2012_609	Adult_2011_22	Adult_2011_26
+0y_2012_618	NA	Adult_2011_7
+0y_2012_620	NA	Adult_2011_14
+0y_2012_622	NA	Adult_2011_6
+0y_2012_623	Adult_2011_52	NA
+0y_2012_629	Adult_2011_52	Adult_2011_6
+0y_2012_634	Adult_2011_52	NA
+0y_2012_641	NA	Adult_2011_45
+0y_2012_647	Adult_2011_37	NA
+0y_2012_649	Adult_2011_53	NA
+0y_2012_650	NA	Adult_2011_9
+0y_2012_653	NA	Adult_2011_7
+0y_2012_654	Adult_2011_53	NA
+0y_2012_655	NA	Adult_2011_6
+0y_2012_657	NA	Adult_2011_7
+0y_2012_661	Adult_2011_53	NA
+0y_2012_662	Adult_2011_53	NA

+0y_2012_668	Adult_2011_53	Adult_2011_42
+0y_2012_669	NA	Adult_2011_6
+0y_2012_671	NA	Adult_2011_40
+0y_2012_673	NA	Adult_2011_26
+0y_2012_674	Adult_2011_53	NA
+0y_2012_675	NA	Adult_2011_40
+0y_2012_676	Adult_2011_29	NA
+0y_2012_677	Adult_2011_53	NA
+0y_2012_679	Adult_2011_53	NA
+0y_2012_680	NA	Adult_2011_18
+0y_2012_681	NA	Adult_2011_40
+0y_2012_682	NA	Adult_2011_40
+0y_2012_683	Adult_2011_37	NA
+0y_2012_685	Adult_2011_53	Adult_2011_15
+0y_2012_687	NA	Adult_2011_40
+0y_2012_689	Adult_2011_22	Adult_2011_26
+0y_2012_690	Adult_2011_53	NA
+0y_2012_691	Adult_2011_53	NA
+0y_2012_695	Adult_2011_37	NA
+0y_2012_698	Adult_2011_37	NA
+0y_2012_699	NA	Adult_2011_9
+0y_2012_700	Adult_2011_53	Adult_2011_9
+0y_2012_702	Adult_2011_53	NA
+0y_2012_704	Adult_2011_53	NA
+0y_2012_705	Adult_2011_53	NA
+0y_2012_706	Adult_2011_53	NA
+0y_2012_707	Adult_2011_53	NA
+0y_2012_710	Adult_2011_53	Adult_2011_15
+0y_2012_712	NA	Adult_2011_18
+0y_2012_718	NA	Adult_2011_49
+0y_2012_722	Adult_2011_53	Adult_2011_15
+0y_2012_731	Adult_2011_22	NA
+0y_2012_734	Adult_2011_53	Adult_2011_9
+0y_2012_735	Adult_2011_53	NA
+0y_2012_736	Adult_2011_53	Adult_2011_15
+0y_2012_769	NA	Adult_2011_40
+0y_2012_771	NA	Adult_2011_36
+0y_2012_779	NA	Adult_2011_36
+0y_2012_781	Adult_2011_39	Adult_2011_18
+0y_2012_783	NA	Adult_2011_18
+0y_2012_785	NA	Adult_2011_12
+0y_2012_795	NA	Adult_2011_36
+0y_2012_796	Adult_2011_28	NA
+0y_2012_797	Adult_2011_28	NA
+0y_2012_799	NA	Adult_2011_36
+0y_2012_800	Adult_2011_28	NA

+0y_2012_801	Adult_2011_39	NA
+0y_2012_802	Adult_2011_37	NA
+0y_2012_803	NA	Adult_2011_36
+0y_2012_804	Adult_2011_37	NA
+0y_2012_807	Adult_2011_39	Adult_2011_40
+0y_2012_808	Adult_2011_28	NA
+0y_2012_811	Adult_2011_39	Adult_2011_18
+0y_2012_815	NA	Adult_2011_36
+0y_2012_821	NA	Adult_2011_12
+0y_2012_746	Adult_2011_53	NA
+0y_2012_747	NA	Adult_2011_54
+0y_2012_748	NA	Adult_2011_15
+0y_2012_749	NA	Adult_2011_36
+0y_2012_750	Adult_2011_37	Adult_2011_17
+0y_2012_754	NA	Adult_2011_40
+0y_2012_765	NA	Adult_2011_12

APPENDIX G: The 167 SNP containing loci included in the study.

Table 1. The 167 SNP containing loci included in the study. The sequence of the SNP containing locus is provided here, together with the SNP ID, the alternative SNP alleles and the primer binding sites. The position of the SNP site in the SNP containing loci sequence is labelled as N. The black line middle of the table marks the two different multiplexes where the primer pairs in the PCR-1 amplifying each sequence were included. This table is modified version of the SET_FINAL object used in the genotyping pipeline.

SNP ID	Forward primer binding site	Reverse primer binding site	SNP containing locus (SNP site marked as N)	SNP alleles
AKAP11_4	CTTGTCTCTTTCCCTAGCCAAT	GATAAATCCTTGTTTCTCTTGTGG	CTTGTCTCTTTCCCTAGCCAATGGCTCTTATCTCCATGGAAACCAGCGTGATNTGGCCTCCCAGACTGTCCCCGCCCACAAGAGAAACAAGGATTTATC	[A/G]
UtagF_SS_147a	CTCCCTCTCCCTCTCTCTCG	CCCTGGAAACTGCTGCTC	CTCCCTCTCCCTCTCTCTCGGTCTATTTCCCTCTCTCTGTAGAAGAAGCTGGGTGTTTACAGTAGGATGCAGCAGGACAGCANGGAGCAGCAGTTTCCAGGG	[T/C]
UtagF_SS_148c	TGTGGGACACAGCACACAC	GGCCTGCTCCACCTCTGT	TGTGGGACACAGCACACACAGTACCACACACAGCAGCCAGAGCCCAGGGATACACAGTGATGTGAANACAGAGGTGGAGCAGGCC	[G/C]
c09_3783_SGT (*six6$_{TOP}$*)	TCTGTTGCTTGTGTTTGTGTGT	CACAAGTGCCAGGCTAGGAG	TCTGTTGCTTGTGTTTGTGTGTTTACTATCATAAAATANCATTTTTGGGCTCCTAGCCTGGCACTTGTG	[T/G]
c25_1441_SAC (*vgll3$_{TOP}$*)	TCTCCTCTGTTGTCATCCAGAA	ACCCAATCAGACCACACAGC	TCTCCTCTGTTGTCATCCAGAATTAATCNGATTGTATTCCTCCAGTACAGAACAGCTGTGTGGTCTGATTGGGT	[A/C]
c25_684F_713R_SACa	AATGCGGTCTTTCATTGCTT	TGCTTGATCTGCTCCTGAGA	AATGCGGTCTTTCATTGCTTATTTTCTTTCCATTCACGTACANTCCTCANGTAGAANGACATGTTCTCCATNAATCCAAGAAACTCTCAGGAGCAGATCAAGCA	[A/C]
c25_684F_713R_SACb	AATGCGGTCTTTCATTGCTT	TGCTTGATCTGCTCCTGAGA	AATGCGGTCTTTCATTGCTTATTTTCTTTCCATTCACGTACANTCCTCANGTAGAANGACATGTTCTCCATNAATCCAAGAAACTCTCAGGAGCAGATCAAGCA	[A/G]
N_208	CCACATCCCTGTCATTGGTA	GCAGATTACACACTAGCCATCG	CCACATCCCTGTCATTGGTANGAGAAAGAGACTGAGATATCTGGGACGTAGGTCGGGGTGCCTTGTAAGTATCCGATGGCTAGTGTGTAATCTGC	[A/G]
N_364	TCATTCTCTGGACTTCACCTCA	AACGGAGAGTCGAACAGAAGTAA	TCATTCTCTGGACTTCACCTCAAATACTTTTTATTAGAGACACTCCTTCTGACATTANTCACCATTCATTTACTGTACTTGCAGTAATGTTTTACTTCTGTTCGACTCTCCGTT	[T/C]
N_473	TCAAACAATTACAGTCCCTGAAGA	TTCATTTGGCAAGAAACATCTC	TCAAACAATTACAGTCCCTGAAGAGGATTTTCTCTCCAATTGGCTTTNAATGAGTCACGTCTAAACTGAGATGTTTCTTGCCAAATGAA	[T/A]
N_864	TGCTACCATCTTTGCCACTTC	TGTGACCCTAACTCGCATTCT	TGCTACCATCTTTGCCACTTCTGAGTAGACAGCAATTCATTGGTTATTAATTCATGACCAGTTGGCTAAATTAAGACATCTNTTAGAATGCGAGTTAGGGTCACA	[T/C]

N_890	GAGCCCAACATCCCAAGTTA	TGCACAGACCCAGAGTTTCA	GAGCCCAACATCCCAAGTTATGANGTTGAAAAATGGAGTGTAGAAAAATAACATGGTTTTGTAGTAGGTCTTGGCTGTTCTTGTTGAAACTCTGGGTCTGTGCA	[T/C]
N_1078	GCAGTAGCTTGTCCCATCTCA	TGTAAAGCCTCTGACCGGAAT	GCAGTAGCTTGTCCCATCTCATAACTGTTCTCAAGAAATGAACTAACGTGTCATTTCTTGAATGGAACAAAGCTTTAGACTATTGTTTTCNAAATTCCGGTCAGAGGCTTTACA	[A/C]
N_1151	CCCGCCCAGAAGTAAAGAGT	GTTGGCTCAATGGCACATAA	CCCGCCCAGAAGTAAAGAGTTGTTGACGTCGCTGGTCNTATTTTGTTTTTCTATTGTCAAGTTATGTGCCATTGAGCCAAC	[C/G]
N_1247	GCGACCAAATTCAAGAGGAA	ACGATCCCATCAATCTCCAG	GCGACCAAATTCAAGAGGAACGAAGCTCTGGACAAGAAGATCTCCCTGTTCAGNGGTGACATCACCAAACTGGAGATTGATGGGATCGT	[T/C]
N_1269	TTTGTGTAAAATTCGGGTTTCTC	AAAAGGGTTAATGCCCCAAG	TTTGTGTAAAATTCGGGTTTCTCNTTAAGTGCAATTTGTTTCCTATCTCCCTATGGTTGCTTGGGGCATTAACCCTTTT	[A/T]
N_1390	GCATTGACACACACATTAGCC	GCCCCTTGTCTTTGTTCCTA	GCATTGACACACACATTAGCCAAGTGGCTAACGTGAGTAGAGTCCCACTAAAACATTACATATGCAGAGTTTTCCNAAGTAGCGTTGGGGGTACACCTTAGGAACAAAGACAAGGGGC	[A/G]
N_1645	AGGCAGTCAAAGAGCACCAT	TGCACAAGTATTCAGCAAACAA	AGGCAGTCAAAGAGCACCATTTTGTGAATTATTCAGACTGNTTTGTATTCCTTGATCATTGTTTGCTGAATACTTGTGCA	[T/C]
N_1852	CCCCTCCGATATGGTATTCA	TGCAACATGGAAAGCATAAGA	CCCCTCCGATATGGTATTCATTCAAAACCCGTAACGCTTTTCNAAGTACAGGTATTTGTTCTATGTTCCATGTTCTTATGCTTTCCATGTTGCA	[T/C]
N_2406	ATCAACCTCCATTAGAAAATGTGAT	AATCTACTTTCAGGCACCGTTT	ATCAACCTCCATTAGAAAATGTGATNTTACAAAGAGAAATGTGCCTTGAATGTCATTTTGGAAAAACGGTGCCTGAAAGTAGATT	[A/T]
N_2488	GACGGTGCCCATGATTAAGT	TCTCTTGAGCCGAGTGTGAA	GACGGTGCCCATGATTAAGTCTTGAATGTAACGTACTTCAAATGTTGGTGTTTATTTTAAGAAATNAAAGGACATTTAATTCACACTCGGCTCAAGAGA	[A/C]
N_2551	TCCCAGTTCAATCAATCAATCA	CGCTGAAGTTGCCTATCAATC	TCCCAGTTCAATCAATCAATCACAAATCAATATCACAAATCAATGATAAGNCAAAATTAACAAAGATCAGATTGATAGGCAACTTCAGCG	[A/T]
N_2597	TCACTTTCAGCTTGATGTATTATCG	AACACTTGACCTGCCTCCAG	TCACTTTCAGCTTGATGTATTATCGTTAATTGTTACTTNATTCGTAAGTAGTGATGCGCTTTTAACAGGTGTAGAAATCAGTATACCTGGAGGCAGGTCAAGTGTT	[T/G]
N_2723	GGATGCCTTAGTTTCCACGTA	TGGGAAAGGTTGATTATGTAACTG	GGATGCCTTAGTTTCCACGTACACTCACAAACAGAAGGTGTAC[C|T]NTCTTTGGTTAGTTATAATTATCTTTGGTACACTTGAAGACAGTTACATAATCAACCTTTCCCA	[T/C]
N_2792	GAGAACCGCTGAATTGATGAG	CAGACATAACCCCACCCAAG	GAGAACCGCTGAATTGATGAGTAANGTCATACTGAACATGAAGTATTATAGATTAAGGCTATAATGGTTTTATTACTTGGGTGGGGTTATGTCTG	[T/C]
N_2222	GGCCACCAATATATACACTAAGTCC	AACCAGTCAACTAATAGTCACTCT	GGCCACCAATATATACACTAAGTCCACGAATACATTTNGCACTCCTAAAATTCATTTTAGAGAATAGATACTTAAATTTTGCTTCT	[A/G]

		GC	CCCTGACTGGCAGAGTGACTATTAGTTG ACTGGTT	
N_2618	CATCATTCCAG GCCACATTT	ACGCAGG ACAATCCC ATTT	CATCATTCCAGGCCACATTTAGTTTGTTC TTTACACTACTTNAGTAAAGCTGTGTGT GAGGTGAACATTCCAGAAGGAAAACAG TCTGCCCCAGTGCAGTCAAATGGGATTG TCCTGCGT	[T/G]
N_2731	CGGCTTCAAAT GACTGACC	CACCGCCT GAATCATT TCTAA	CGGCTTCAAATGACTGACCCATGACCCA GACATAGAAGGCTGCCATTTTAATATTTC TAGTTTAGTTTGGCTTGTATCCANGTCCA TCTTAGAAATGATTCAGGCGGTG	[A/T]
1_1	TATCATGTACA GGTTACCCCATT G	AAGACCTG TCATTTTG TGACCA	TATCATGTACAGGTTACCCCATTGGGATT AAAAATNTATTTTACAAGAGAGACTGGT CACAAAATGACAGGTCTTT	[A/C]
2_1	ATGGTAACAGG TGTCGTCGTC	CTGTAGCT CTGACTGC AGTGGT	TGGTAACAGGTGTCGTCGTCGAACAAGC AGGGATCAGGCAGCACCTGNTGTAGCA GCCCAGAGTGGAGCAACACCTCAG	[A/G]
7_1	TGGGTGGGGAC ATAATACATTT	GGACGTCT CAGAGCG ATCA	AATCATGGGTGGGGACATAATACATTTT TTCATGTTTTCTATCAAGTCAAAAATCTA AAACGAAACACTTTTTTTTAACTACTCCA AACAGCCTANCTGATCGCTCTGAGACGT CCG	[T/G]
8_1	AGTGGGCAGAA ATTGGAAACT	TGATCACT GAGGATTT GAGTGC	AGTGAGTGGGCAGAAATTGGAAACTGC TNAGGAGTTGTGTGCAGTCCCCACTCAT GGGGAGAGAAGGGGGAAGCACAG	[A/G]
10_1	TGGATTTAACCT TTCGAGTTTCA	TGAACTGT ATGGCGG TGTCTAT	TGGATTTAACCTTTCGAGTTTCACACACT TTGTTGTACTTCTAACCAACATCTANCTA TAGACACCGCCATACAGTTCAAA	[C/G]
11_1	GACAAATTGCC ATTAACATTGC	CTTTGGAA AACATGG GACAGTT	TGCCATTAACATTGCTTTGCAAGATCAGT AACACACTTTGATCAAGATTACAAGNCA GAACTGTCCCATGTTTTCCAAAGGT	[T/A]
12_1	CAGCAAAATCA AATGGTTTACA GT	GATTACAT GGCCATTT TCTTTGA	CAGCAAAATCAAATGGTTTACAGTTTTC AATGAGCTACAATTAATAAAGGACTGGG TGTANTCAAAGAAAATGGCCATGTAATC AATGATCCAACTCCTTGTGACTTGGCAA CCA	[T/G]
14_1	AGATGTGTTGC TGGGAGAGATT	ATTGGGTG GATAATTT GAGCAC	TGTTGCTGGGAGAGATTTGTACAAGTTG TATATNACCATCTGTATGTGGAATTTGAC ATGTCTTAATTCAGTGCTCAAATTATCCA CCCAAT	[A/C]
15_1	CACTCCTATCCA TACAGCATTTTT C	TTATACTG TTGGCAGT CAGTTGG	TCATGAAATTCTATGTTGCCTACAAAGTN CCAACTGACTGCCAACAGTATAATTTGA TGCATAACTTCTATCATGTCTCAGAAGG GATTGCCACA	[A/G]
17_1	TTCTTAGCGTG GTCAGACTGTT	TCATGTGA AGGACCA GCTAAAC	GCGTGGTCAGACTGTTACTGGCCGCTTG GAGGAGGGGGGTTGTCAGGGTAACGG GNGGTGTTTAGCTGGTCCTTCACA	[A/G]
18_1	AAGAAGGAGAT AGTTGGGCTGA	CTCCAATT GGTTGTTT TTGTTG	AAGGAGATAGTTGGGCTGAGGGGGGGT ACAAAACAACATATTTTACCTCAAAAGC AGTNACTTGTCACTCATATTTCAAAACAA CAAAAACAACCAATTGGAG	[A/G]
19_1	TGTAGCACCAT CGTATACTCTTC TGT	CACTGGCT TGATGACT CCTGTA	AGCACCATCGTATACTCTTCTGTAATAAT TGTGAAGTAACATGTTATAGATTTCTCAT TTCTAANTTGATATGTGCTTTACTTGNAC AGAATGTTACAGGAGTCATCAAGCCAGT GA	[T/C]

21_1	TTCTTGTGTAGGCTGCTTTCCT	GGTCACGTCGCCAGAGTTAG	TTCTTGTGTAGGCTGCTTTCCTATTTGAAATAAAATCATANGAGAGAAACAAATTGGCATGTTAGCTAACTCTGGCGACGTGACCG	[A/G]
24_1	AGTGCATGACAATGAGCCAGT	AAATAAATCTTTCAGTGGCACACA	AGCTGAGCAGTGCATGACAATGAGCCAGTTCCAAGNTAGGCAGNNNAGACGGAGTAGCCTCTGTGTGCCACTGAAAGA	[A/G]
26_1	GCCAACAACCTTCATGTCCTA	GGGAAACCCGTCATTTACAAC	GCCAACAACCTTCATGTCCTAAGAACTGACATTTAAAAAGTGTTTCAATGAATAACCTGCNACAAGTAAATATAACGTTGTAAATGACGGGTTTCCCG	[A/G]
28_1	TGGACCTTTGGTGTAATACTCG	CATTCAAGTTAAACCCTTCAGCA	TGTTCATTACTTTTGCGCATTTCTAAATGGACCTTTGGTGTAATACTCGGTTGNCCGTCCTGGTCTTGGAGAGCCTCAGTG	[T/G]
29_1	ATACCCCTTTGGGTCACCTTAC	GGGGTGGTACGGTTGTAACTAT	ACCCCTTTGGGTCACCTTACCCNACTCACCCCTATACCAATTGTGAGGAGTTTGTATAATAGTTACAACCGTACCACCCCAATG	[T/C]
30_1	GGACCTTTCGGAAGAGAGAAGT	GAGAGGAGTAGGGGTCAGATTG	TGATTTCTAGGGGACCTTTCGGAAGAGAGAAGTTACAACCAGTCNGAAACACGAATATGAATGGGAGGGACAATCTGAC	[A/G]
31_1	GCCCTGCAGTATTGAGAAGGTA	CCAGTGTCGATTGTTTGATGTT	GCCCTGCAGTATTGAGAAGGTAAAAATCCCATTGTTACCGAAAGACTCTTGCTNCCAAAAACTTGAACATCAAACAATCGACACTGGAA	[A/G]
33_1	GATACAAGTTCTGCTGCAAAGG	ATCGATCATGTCCTTCCTCAGT	GATACAAGTTCTGCTGCAAAGGNGAGTGACATTCTAATTGGTATTAGTTGACATGGATAGATACTAAATCACTGAGGAAGGACATGATCGAT	[A/G]
34_1	GAAATAAACATTTTGCCATGGAT	TTTAGATTTCCCGCAACAGACT	TGCCATGGATTAGTAAATGAAAAATCAAGTTAACATTGAGACAGATTATTAAAATAGTCTCNTTTTATGGTGGAGTCTGTTGCGG	[A/T]
35_1	ATAGCCAGCACCATATTGCCTA	AGGTCCTTTTAATGCCCATAGC	TAGCCAGCACCATATTGCCTATTACTGGTGATATTGAAAGCTCTATCAGTCNGCTATGGGCATTAAAAGGACCTTT	[A/G]
38_1	CTCCTCTCTCTGGGGAGAAAC	TCTCATACATCCCATCATGCTT	GTACTGGGGCCTCCTCTCTCTGGGGAGAAACNGGACTGTCAACGAGGTGGTCATCACCAGGCTGAGTGCAAAGCATGATGGG	[A/C]
39_1	TAACTTTCCCCCAAAAGTTCCT	CCACTGAAATTGTTGGAAAAGG	CTTTCCCCCAAAAGTTCCTCCAAACATACCAAGTTCTATTTGACAANAACCCTTTTCCAACAATTTCAGTGGCCCAG	[A/C]
47_1	TCTGAGATGCTACTCATTGTCCTC	GCCTACAATCAGGTGTCAGTTG	CCTTCATAGTGGCGCTCAGCAGCACCTCTGAGATGCTACTCATTGTCCTCGGCCATTTNGCCCATCGAAAGTTCTGTGCCCCG	[T/C]
48_1	CAAGGAAACGTTGGCATTAGAC	CGCAACGACTTCTGTGGTAGTA	GACAAGGAAACGTTGGCATTAGACTACTTTTGAGGTCTTTGATATTGGNGTGAACTACTACCACAGAAGTCGTTGCGGACATAC	[T/C]
49_1	AGGCACGCATATCTTTAGCAAG	ACGTAAAACATCTGCCTTCTGC	GCAGCACCACAAGTTAGCAACTTGTCAATGATGGAGGCACGCATATCTTTAGCAAGACTGTAAAATGTCGANCAGAAAGTAAAATGTCGCCGGAG	[T/C]
50_1	TAGCCTCCAACAACAAGTCTGA	AAATTAATCGGAGCGACACTTG	GCCTCCAACAACAAGTCTGAGTTTTTTTCGCCGCGAGNA.*GTGCTAGCACGCTGGGAGGCGGACAAGTGTCGCTCCGATTAATTT	[T/G]

TN_236	GGGTTGAGTAGGGCTCACAG	CCACCTTCACTCTGTCGTTG	GGGTTGAGTAGGGCTCACAGTTCAGAACGTAGTTAGCAGACATTGAATGAAATGTAGTCAGGTACTAGNTCCAACTATTAATGAAAGGAAGTCACACAACGACAGAGTGAAGGTGG	[G/C]
TN_399	GAACCAAAGGGAAAGTGTTCA	GCATTCCATCGGTCCATT	GAACCAAAGGGAAAGTGTTCAGAACACTGGAACAGATCATTGTCAGTAGAAGGTCAAAAAAGNACATTAACAAATTCTAAATCAAATGGACCGATGGAATGC	[T/G]
TN_423	ATCCATCTACCGCTCACTTCTT	CCAACATGCCCTTGGAGA	ATCCATCTACCGCTCACTTCTTCTCGATGCGATCATCTGTTTTTACCACCCCTGTCCCAGANTCGCTCGCTCACGCTCTCCAAGGGCATGTTGG	[A/G]
TN_489	AAGAAACCCGAGCCCATTT	TGTAAGGGATTGCTGTATTGGA	AAGAAACCCGAGCCCATTTCAAGGCCGGCACGGAAGAGTAGTATCANGCTGTCTGGCATTGATTTCCAATACAGCAATCCCTTACA	[T/C]
TN_1026	TCAAGGTCAATGCGTCTCAC	CGGTGGATGCCTAAGTCG	TCAAGGTCAATGCGTCTCACTGTACCTTACATTAAGTGTTGTGCAACAAATATCCTGAGCTCCNGGAAGCCCCACCACTCAGTCACGACTTAGGCATCCACCG	[A/G]
TN_1499	CATTTAACCCTTAGCTGGTATGC	GGGAAGAAGCCCTCGAAGT	CATTTAACCCTTAGCTGGTATGCTTTATTAATAGAGAAATAATCATGTACATAGGCCTACAGAGTAATGGTTGTATCGAAGGTTGNTTAGATGTGACTGAACTTCGAGGGCTTCTTCCC	[C/G]
TN_1510	AAACTTACCGGACGTGGAAA	GCAATTCCGAGATGGTTGAC	AAACTTACCGGACGTGGAAAATGTTTGAACAGACAGATGCNGCGTTTAAAAACAACTCGAGAAAATAAAACATCACAAAAAACAAACACATTGAAATGTCAACCATCTCGGAATTGC	[A/C]
TN_1544	TGTATGTGTGCCATGCTATGTG	TTGTGTCAATGTTGCTACTGTTTC	TGTATGTGTGCCATGCTATGTGACATGATCAGGCTAGTNTTACCTGTTCAGGTAGACAAAGAAACAGTAGCAACATTGACACAA	[T/G]
TN_1576	TGCCAGTAAGGGTCAGAGGT	CAGGGCAGTGTGAGTGGAC	TGCCAGTAAGGGTCAGAGGTCACTACCCNGCAAGCCGAGGGGCGAGGAGGTTGCTGGTGTCACAGCCCTGCTGGAGTCCACTCACACTGCCCTG	[T/A]
TN_1734	CCCAGTGAGGTTGGTCAGTAG	TCCAAGGGTGTGAGTGAGTG	CCCAGTGAGGTTGGTCAGTAGAGCATCTTACTACNCATCTTATTCACTCTTACACACACACATTCACTCACTCACACCCTTGGA	[T/C]
TN_1773	AGAAGTTGACCGGACGGAAT	CCCTCTCCCTCACCCATC	AGAAGTTGACCGGACGGAATCCGAACCAAGTANCGGTCCCAATAAATATTTGAGCGGAGTGGGGAACGCCATGATGGGTGAGGGAGAGGG	[A/T]
TN_1937	GGGAAGTTTGTGGTTGGCTA	TCAGTGTGACTAGAGCCCAAATAA	GGGAAGTTTGTGGTTGGCTAGCGCTTGAGGCTGTGCGGTGTTTATTGTTTGGATTTCAGTGAGGCTAGCTAGCTCGATATGGCNCGTGTTTATTATTTGGGCTCTAGTCACACTGA	[A/G]
TN_2066	AGCCCAGTCAGGCAGCTA	AATCCGTCAATCATGCCAAC	AGCCCAGTCAGGCAGCTANTGACAGACAGCAGTGACCCTAAAGTGAGTGAACACATGTTGGCATGATTGACGGATT	[T/A]
TN_2333	GGCTACCTCCACTTTGGTGA	CCATGTTCCTGTTTATGTGACG	GGCTACCTCCACTTTGGTGANGACATCAAAGGAAGCATAACCCTTAACCTTCCTCACAACGCATGGTAAGCCCCGTCACATAAACAGGAACATGG	[A/G]

TN_2606	TTCACAACACAAACAGGGAATTA	TCATCCATTTATCCTTGGCTTT	TTCACAACACAAACAGGGAATTATCTTCAAAGCCACNGTTTTTTCTCACTTGGCTAAGGCCCTTTCATGCTGAGGGGAATCAAATTGTATTGCGTTACAAAGCCAAGGATAAATGGATGA	[T/G]
TN_2787	AGAAGGTGAAGCCCAAAGC	TTAAGTAATCTGCCCAACCATGT	AGAAGGTGAAGCCCAAAGCTTTCGTCTTCTGATTAATGTGCTTCTTTTGCAGGAAGTATTTCAGAGACTCACATGGTCAAACAGCTNTAAAAGAAACATGGTTGGGCAGATTACTTAA	[A/G]
TN_208	GAACAGGACAGGGAGAGACG	CGTCCGAAGAATATAGGAGTGTG	GAACAGGACAGGGAGAGACGTCCCTTGTTTCAGGTGATAGCTTAATTCTNTCTAAGCAGTCCTTTACACACTCCTATATTCTTCGGACG	[G/C]
TN_443	AACAAGGATAGAAATTGGAAGTCTG	GGCAGCAGAGGTCTCACG	AACAAGGATAGAAATTGGAAGTCTGACAGTNATGCCTTTATTATTTTTAGTCCTATCACAACCTGACCGTGAGACCTCTGCTGCC	[T/G]
TN_473	TCCTCCCATCCTAGAGAGACC	CTGTGTTCTCAGCTACCTCTGTG	TCCTCCCATCCTAGAGAGACCACTGGAGATCGACTGTCCTCCCAACCTAGAGNGACCACTGGAGATGGCTAACAGTCTCAACACAGAGGTAGCTGAGAACACAG	[T/C]
TN_890	AAGAGAGTGCCCAGAAATCG	TTGCATGGGTTTAGAGGTGA	AAGAGAGTGCCCAGAAATCGGATTAGATTGGCAGAGTGAATTTACNAACGCTCCCTAGACTCTCACCTCTAAACCCATGCAA	[A/G]
TN_1078	ATCCTTGCTGCTCCATGC	CATTTAGCGAGTCCTGTATAGTGAAA	ATCCTTGCTGCTCCATGCNGAGGTTCCTCTCTCACGGGTAGCTGTTCCTGTCTTTTCACTATACAGGACTCGCTAAATG	[T/C]
TN_1151	CCATATTACCAGAGCTTCATCTTATTC	TCACCATCTATTGTTACCTCGAAA	CCATATTACCAGAGCTTCATCTTATTCTTNATAAGCCTTTCTATGGCGGATTCGCAGTCACAATTTATGGAAGATGCGAAAACTTTCGAGGTAACAATAGATGGTGA	[A/T]
TN_1212	GCCATTCCAGTCAGTCACAA	GGAACAGCTCTCCTTCATTCA	GCCATTCCAGTCAGTCACAAACACACTTCTAAGCAGNGGCACCCATCTTTAAGCACTCACAGGCCATTTATTTTTACTTCCTTACACAATTCAAAGGGATGAATGAAGGAGAGCTGTTCC	[T/C]
TN_2056	GGAGGATAGGATACCGATAGCA	TTGATGTAGGCAGGTCTGTAGG	GGAGGATAGGATACCGATAGCAAAAAGCAAGAAATNACATCATACAGAAGTTTATAAGCCACCTACAGACCTGCCTACATCAA	[A/C]
TN_2268	ACAACTGGGCACACAGGAG	GTCCAAGGGCTGTCACCA	ACAACTGGGCACACAGGAGCCATGCTGACGTCTATCAAGTTGCCTATTGTGATCAGAGCCGAAATCAATTTGGTTTTGGTGNTGGTGACAGCCCTTGGAC	[G/C]
TN_2350	GGCATGTGCTCAACAACAAA	CAGATTGTGCTCCTTTCCTTCT	GGCATGTGCTCAACAACAAAAAAAGGCACGCCCAAATAAAATCCCTNCAACTGGGTCACAGGCTAGTTGAGCAATTATTACAAGAAGGAAAGGAGCACAATCTG	[A/G]
TN_2536	CTATTGCGTTGACATGCACA	TACATGCAGCTCTCGCTTTG	CTATTGCGTTGACATGCACAACTCAGACTNGTGCGGCATAACCAAGATGCACTTGTTTTGAGATCAAAGCGAGAGCTGCATGTA	[T/G]
TN_2736	CGAAACGGCAACAAAGACA	TGCACAAGCTCAGTATCCATTT	CGAAACGGCAACAAAGACAAAAGTTCCCATGTACTGTAAGTAGGGTTAAAATAGATTTNAAATGGATACTGAGCTTGTGCA	[G/C]

TN_2792	GCGTGTGTTTGAGCATGAGT	GGACAAGTCAATATGGCTCTTCA	GCGTGTGTTTGAGCATGAGTCAGTGTACATTTGAGACTTTGTTTCTGGGACTTTAGTGATTCTCNAGTATTACAGTTTCTGAAGAGCCATATTGACTTGTCC	[A/C]
TN_2841	GGCAGTCTTTCATATTTCTACAAGG	TGCCACAAGGAAGTGAAGG	GGCAGTCTTTCATATTTCTACAAGGCCCCTGGGTTCCCCCACCCCATGAGGCATATACACTGACCNATTCCTATGTCCAAATGCATCCTTCACTTCCTTGTGGCA	[T/G]
TN_1195	TTTGTTCATGTGCTGTGCAA	GCCCTTGATACTGCCTTGTG	TTTGTTCATGTGCTGTGCAAAACCCGTACCCCTAATGGCCTGCTGCCCTGGACTATTCGAATGGGCCAACGTCACAAAGTTTNTCACAAGGCAGTATCAAGGGC	[T/G]
TN_2603	TGAAATGTTGTTGTGGTCTTGG	GTGCAGTGACATGGGATCAG	TGAAATGTTGTTGTGGTCTTGGTATTATGGNCCCACCCTACATGACCCATGTGTATTGGGTCATATCATCATCGCTGGCCTGATCCCATGTCACTGCAC	[C/G]
TN_1642	CAATTAGTGCAGATGGGTAAGGT	TCCGTTCTTTGGCTCATTTC	CAATTAGTGCAGATGGGTAAGGTAATGAAAAAGNCCAATAATTACCATGATTTCTTTTGCAAACATTTTGTCATTTGAAATGAGCCAAAGAACGGA	[C/G]
TN_645	CAGGGTGTTGTGTAAAGTTGGA	GTCAAGCACATGGTTAATGTCAG	CAGGGTGTTGTGTAAAGTTGGAATGTATGAGGTGGAAAAAAAACATTCTAAGTCNGTGTGTCTGACATTAACCATGTGCTTGAC	[A/G]
TN_1307	AGGTCAGCCATCTGAACCA	ACACTTTCACACCACATAAACCA	AGGTCAGCCATCTGAACCANTACTGCCCTCCAGTGGAGGATTACGCATGGGAGTGAATGGAGCTGAAGTCAAGTGACTATGGGCTTCTATTATTATGGTTTATGTGGTGTGAAAGTGT	[T/C]
TN_2594	ATCCCTGTGTTGACCTTTCTG	GATTCTGGCGAGTGGGAAT	ATCCCTGTGTTGACCTTTCTGAGAGACCACAGNAACGCTCAGTTCCGAAACATGATTGACCTTACTGCAGTTGACATTCCCACTCGCCAGAATC	[A/G]
TN_1370	TGTATAAAGTGAGCTGAAGAACTGG	GGGTGTTCCAAAGGCTGA	TGTATAAAGTGAGCTGAAGAACTGGCCATAGAGGGGCCAGAGTGAAGAATAATNTTCCTCAGCCTTTGGAACACCC	[A/G]
TN_2240	TCTTTGGTCTCTGTCCCAGTC	ACATACCTTTCGCGTTCCAC	TCTTTGGTCTCTGTCCCAGTCTANTCCAATGGGCCTTTCCTGCTCTTTGTTTATAGGGGACAAGAGTGGAACGCGAAAGGTATGT	[C/G]
TN_376	CCATTTCATCAGAGCGATCC	TCAGGTCAGTTCCAGTGTGC	CCATTTCATCAGAGCGATCCTGGCACTGTGCCTTCCCATCACACACCTGCTGCTGGTCGATACACTTCNTCCCATGAGCACACTGGAACTGACCTGA	[A/T]
TN_1263	AATACTGTTTACCTGTTCCACCAAA	TCCCAGGACAATGGACTAACA	AATACTGTTTACCTGTTCCACCAAAGCCTGGANTTAATCTAGCTGTTGCTAGGGTGTTAGTCCATTGTCCTGGGA	[A/G]
TN_2431	AAGACCATTTGATTGGCATGT	TCTCTATAATGTTTGTACTGGGACGA	AAGACCATTTGATTGGCATGTATGTCAAGGGNTTCTTCTAATTATGCATGACCTATTGGGGCGTCGTCCCAGTACAAACATTATAGAGA	[A/C]
TN_921	ACTCACCTGCCATCCTGTTC	AGATTCCATGCTCAAGTGTAAGG	ACTCACCTGCCATCCTGTTCAGGTGTGACATCACCCAGGATGCATCTCAGTACTCAAAATAGCTTCCTATCTCCTTTNATTTATCCTTACACTTGAGCATGGAATCT	[A/C]
TN_1912	TTCTACATGCGCCATTCTCA	CGATGACTCCGCCTACCTAGT	TTCTACATGCGCCATTCTCAGGCCCTGGCCCCTCCCCNTTCCATCACTGGAGCTCTGGTTGGACTAGAGAAGACTAGGTAGGCGG	[A/T]

			AGTCATCG	
TN_2692	ATGTGGCAAGAATGCTCCA	GCTCTAGCTCATTTGTGATGATTG	ATGTGGCAAGAATGCTCCATGGAAGGGTACAAAATATTGTGGACAAAAAAACAAAAANAAAGACAATCATCACAAATGAGCTAGAGC	[T/C]
TN_371	CCTCGCTGTGACCCTGTTA	CCTTCCCGTTACCAGTATGC	CCTCGCTGTGACCCTGTTACTNAAGTATTTTCCATCTTGGCACACTCGCTTTCTGTTCTCAGGAAATTGTGCTGCAACACCCCTCCCCCCAAACCAAACAGCATACTGGTAACGGGAAGG	[A/C]
TN_2223	AGTGCATGTCCTCTCACCTGT	TCTCTATGGATCGTTGTTTCTCTG	AGTGCATGTCCTCTCACCTGTACAAAAAGCATGTTCGCAAAACNGCCAGAGCTGCATACAAATATTTTCAGAGAAACAACGATCCATAGAGA	[A/G]
TN_2511	GAGCCGTCCTACTCCCTCA	ACGGTGAAGATGACCACTCC	GAGCCGTCCTACTCCCTCATCCCTGTCCTCTACATGCTCATCTTCATCCTGGGCCTGTCTGGNAATGGAGTGGTCATCTTCACCGT	[A/G]
TN_765	GTCAACGCATCAGCAGACAC	CAACCTCCTGCACACATCAC	GTCAACGCATCAGCAGACACCAGTTTCAAGATGGCCACCACCGTAGTNCAGCTCTAATGTGATGTGTGCAGGAGGTTG	[T/C]
TN_1122	TTAGGCCATGCAGTATCCAA	CGACTACTATTGACCGACAGCA	TTAGGCCATGCAGTATCCAATGTATAGTTTGTGCGTTCTGTGAAAGAGTGTAATGGAGGTNTAATTGTCTGAGGGCTCATGCTGTCGGTCAATAGTAGTCG	[A/G]
TN_2455	AGAGGCATGTGTTATTGAGTCG	CCAAGTCAACCGCATTGTAA	AGAGGCATGTGTTATTGAGTCGCNGTGTGGAATGTATTGATGTTTTTGGCAAACTGTTTTCTATTCCGGTTATTACAATGCGGTTGACTTGG	[A/G]
TN_1488	GCCCAGGTCTGAATGTCTGT	GGCACTGGTTGTGTGTCTTC	GCCCAGGTCTGAATGTCTGTGTCCATTTCTCTGTGGATGGTGAATAGATGTCAGAGTNCAGACGGAAGACACACAACCAGTGCC	[C/G]
TN_1294	GTGTTTCTGTTGCTGCTTGG	TTTCAGTGGTGAGGCTGATG	GTGTTTCTGTTGCTGCTTGGGTGAGCATGTGTGCTGCNTGACCTCTGCTGTTAAACACATCAGCCTCACCACTGAAA	[A/C]
TN_2758	CAGTGAGGAGGATAGGATTCAGTT	GGGCAGGGTCTGTAGTTCTCT	CAGTGAGGAGGATAGGATTCAGTTCCTGGTNATCTGTGGCGTACGCTTCCTGTCCCAGTGTGTCAGTACTGAGGCACAGACAGAGAACTACAGACCCTGCCC	[T/G]
TN_2293	CCAGACACCTCTTCCTCCAG	GCCTAGACCAGACCCATCCT	CCAGACACCTCTTCCTCCAGGGCCAGTGGTGGAGCAGTCTCCTGGGCCTGGGTGAGTCCAGTGATGGNATAGGATGGGTCTGGTCTAGGC	[T/C]
TN_861	CGTCATACCTGTGGCTGATG	GGATCTGTGACTGTGCAAGG	CGTCATACCTGTGGCTGATGCACTGTAACTCCTTACCCNTCCATCTTTATTTACACCAATGATCTATGATTCTTTCCTTGCACAGTCACAGATCC	[C/G]
TN_1339	GGTGTTCAAGGCTCCCATC	ATAGGGCTGACGGCTGTTC	GGTGTTCAAGGCTCCCATCCGCCCGGACATTGTCAACTTTGTNCACACCAACATGTGCAAGAACAGCCGTCAGCCCTAT	[T/C]
TN_1071	TTAACTCATCCCGCTTCGTC	GTGTTATGCGTGTGCCATTC	TTAACTCATCCCGCTTCGTCNTTTGTCATTTGTCAATAACTCAAAGGTGATCATTTACTCCATGACGAATGGCACACGCATAACAC	[T/G]
TN_2753	TCGGAAGGAGAGTGATCTGAA	CAAGGCAGTGTTTGATTTGC	TCGGAAGGAGAGTGATCTGAAAAGTGAGGCCTTGATATCTTACANAGTAAAGACTGTCAGTGTGCTATAAGTTCGAAACTAAT	[T/C]

			TTGTCATGCAAATCAAACACTGCCTTG	
TN_2510	CCCTAACACACT TGCTGCTG	CAAAGAT GAATCCTA CCACTCAA A	CCCTAACACACTTGCTGCTGGGACATGT TAACAAACCTACAGTGCATAGCTTTGAG AAATTCTGAATNATCCTGGTCAATGATT AGTTGTTCTTTTGAGTGGTAGGATTCATC TTTG	[T/G]
TN_1881	CTTTACTGAGG GCGATGAGC	TCCGCTCT TTCTTCCT GTCT	CTTTACTGAGGGCGATGAGCTGTTCTCA TTACACCACTCACTGAGAGAGTAAAGTC ACGGTGTAATATTAGCTCTTCTATTGTTC CTGAGCCCATCTNAAAGACAGGAAGAA AGAGCGGA	[T/G]
TN_1153	GAGAATACCAC TTATGCCTCCTC T	GAGTAAC GCACGCC GACT	GAGAATACCACTTATGCCTCCTCTAGACT CGAGACCTCCATGCATTTAGACCCAGAC TGGTCTTTCTCCAAGCTGCTTTCTCTGTC AGTCAGTGGGNGAGTCGGCGTGCGTTA CTC	[A/T]
TN_390	GCGAGGTTGAC CACTCTGTAA	GGCACATT CTGGGAC AGG	GCGAGGTTGACCACTCTGTAAGCTCGCT CCCANCCAGTAAGAAGGGGAGTGGTGA CTCTGCCTGCTCTGGTTCCAGTGGCTCAT CACCTGCCCTGCCTGTCCCAGAATGTGC C	[A/T]
TN_2304	CTCTAGCCTGCT CCACATCC	GGCTGAA CAATTCCT CCTCA	CTCTAGCCTGCTCCACATCCGCACTCTCC CCTGAAAACACCCACAGGTGAAGAGAA CACACACACACAACTAACCTGCAACAAC ATACACANGTGAGGAGGAATTGTTCAG CC	[A/G]
TN_2502	TCTTTCGGTCAA CATGGACTT	TCCTACAG GGACAGTT ACGACA	TCTTTCGGTCAACATGGACTTTTCTTGAA GAGTCAGCCAAATGATCTTGAATCGGAG TGTTACTCGTTNACAGCATAGCTAGCTG TCGTAACTGTCCCTGTAGGA	[T/C]
TN_1468	ACATTGGTTCCA CACTGATGTC	CAAGTAGT TAAGGGTT CCTCTTTC A	ACATTGGTTCCACACTGATGTCATTTCAA TGTTGTGATGTTGAACAGATTTACTGTTC AACATGTGNAGAGGAAAGTGAATTATG AATAGGTTCTGAAAGAGGAACCCTTAAC TACTTG	[T/C]
TN_733	AACACAATTCCC GTGGATCT	AAGGATTT CGGGCAC ATTTA	AACACAATTCCCGTGGATCTAATGTAAA ATGTGACGCAATCTCNCATGAGAGATTT GTAAAGCCTCCAGTATGCGTTACCAAAT GTAAATGTGCCCGAAATCCTT	[T/G]
TN_347	TGACGAGACCC AAACAGACA	TGTAATTG AGTTGCGC TGATG	TGACGAGACCCAAACAGACANGTGAAC ATGTCACCCTGTGTGTCACAATATCCCCA CGTCACACCCAACCAAATCACTGATACT GTCTGCAACATCAGCGCAACTCAATTAC A	[A/G]
TN_1744	ATTTACGGACA CGCACACAA	ACTCACCC ACTCCAGC TCAA	ATTTACGGACACGCACACAAACACAAAC GGGTTACCTCCCTATTACNTTTGCGTAAA CCAATGTGTTTTTGCAGTGTTGGACACA AGCTTTTGAGCTGGAGTGGGTGAGT	[T/C]
TN_680	CGGATTGAGCA GGCTTTC	AGCGTTAG ACCGAGA GAAACA	CGGATTGAGCAGGCTTTCCCCCATCAGC GGTAGTCACTACN[A|T]GCTTGACAACT GCCTGTTTCTCTCGGTCTAACGCT	[T/G]
TN_1385	TCTCAAAGTAG CAGCAGACACC	TTGGACTT CTTCTCGA TCAGG	TCTCAAAGTAGCAGCAGACACCCTGGCN GTCAGACAGAAGAGGCGCATCTATGAC ATCACCAATGTACTGGAGGGCATCGGCC TGATCGAGAAGAAGTCCAA	[A/G]

TN_343	CCAGTTCGGTCTCTGTGTGA	AATCAGTTAATGCGCCCAAC	CCAGTTCGGTCTCTGTGTGATTCATTCATAATGTAATCAAAAAGCATTGCTCTTCCTTCATTAGGAATAGAATGGTCCAACGACAGTAGATAANATGTTGGGCGCATTAACTGATT	[A/C]
TN_220	TGTGTTGGAATGTGATTGCTT	TTCGTTTATTGATCTGTGTGCTG	TGTGTTGGAATGTGATTGCTTTTTCAACTTAAAAATGTATAAAAATAAAGTTAANGTTTTTGTTGCTATCTAGCAGCACACAGATCAATAAACGAA	[A/G]
TN_2672	AACCATATTAGGCAGGGTTGC	GATGCAAGGAATGTACGGAAA	AACCATATTAGGCAGGGTTGCGTATCAGTTGCTGCTAGAAGCGTGAAACAGTNACGGCTCTGTTTCCGTACATTCCTTGCATC	[A/C]
TN_2128	CCTGAGAAGAACGCACAGGT	CAATTCTGGCCTATACCTCCA	CCTGAGAAGAACGCACAGGTGTAGTCATTATGAATGATTATTTTTATTTTTTTTATGAAGTGGAGATACTATCATATTTCGTACACNTATTGTGGAGGTATAGGCCAGAATTG	[A/G]
TN_1235	CACACACATGCTCCTTGTAAATG	ACCGGCTCTTCCTGTAAACTAA	CACACACATGCTCCTTGTAAATGTTTTAAAAAGGTCAGNGTGGAGTCCTGTCTCTTACAAGATCAAGTATTCAAATTGTTTATCCGTATCGATTTGCATTAGTTTACAGGAAGAGCCGGT	[A/G]
TN_1579	GGTAAGCCCACTTTGCAGTC	ACATCCACACCCGAAACATT	GGTAAGCCCACTTTGCAGTCTACATTCAACACTGCATTTANCAGGGTGGATTTGGTGGAAGTAATGTTTCGGGTGTGGATGT	[A/G]
TN_2846	CAGATCATGCAAACTACCAATCA	TTCATTCCTGGTCAATTCTCG	CAGATCATGCAAACTACCAATCAAAAGTTTGGGGATATCCTGAANGAGCTGATTGATATGACCACTCGAGAATTGACCAGGAATGAA	[A/G]
TN_1939	CTTGCCCAGATCACAATCAA	TGAGTGGAAATAAGCAGAGTGAA	CTTGCCCAGATCACAATCAACTCTGCTCCTGNTGACAACCAACCCCAGTCCTGAACTCAGCTAATTTTAGTTTTCACTCTGCTTATTTCCACTCA	[A/G]
TN_1024	ACTGGCATGTGTCTCCCTCT	CGCTGGTGCTGATAGAGTTG	ACTGGCATGTGTCTCCCTCTGGCGTCGTCACGACAACCTTGGACGGGAACCAGGCGTCCTGGGGGAAGAGTGGGAGNGGCTGTTTGTCCAACTCTATCAGCACCAGCG	[A/C]
TN_917	CCAGGGTGAAATTGGATAAATG	AATTGAGCTGTCTGTTTCTGAGC	CCAGGGTGAAATTGGATAAATGTCAGAATCTGAAGTNAAATTGGTGAGATCTTGTTGGCTCAGAAACAGACAGCTCAATT	[G/C]
TN_1735	TATTCCATTGACAGCCACGA	AACATGACGATAGCGATGAGC	TATTCCATTGACAGCCACGAGCACGCTCTCCCCGGGTTCTACCGTGTTGAAGACNGCGCACTCCATGGCTGCGTGTCCGGAGCCGCTCATCGCTATCGTCATGTT	[T/G]
TN_2494	ACTTGAACAGGTTGCCGTTT	GAGTTCGATGGAGGTTACGC	ACTTGAACAGGTTGCCGTTTGAGACNCAGCGGTCTCGTTGGTATCACGTCACATAAAGTCCTCGTAGTCTTGTGCGTAACCTCCATCGAACTC	[A/T]
TN_1514	CAGGACAGCGATTACTCAACC	TCGGGCGTCTGAAAGAAAT	CAGGACAGCGATTACTCAACCAGTACTGGAGGAATACTTTCTCTTTAATGAGTCNGACGGGAAGGATTTCTTTCAGACGCCCGA	[A/G]
TN_2417	TGGCTTCCCATGTATAATTGC	CCTTATTGTTTCCATCCGTCA	TGGCTTCCCATGTATAATTGCTCCTAN[G|T]CCAGTCCACACGCTTCATCCTCCTACACCAGGACTGTGTAAAGAACTGCATGACAGCCATGACGGATGGAAACAATAAGG	[T/G]
TN_1166	CGGGCATGGTAGTGTCAAA	TGCTACAGAGAGATAGAAAGAC	CGGGCATGGTAGTGTCAAAAGGGCACAACAGTATCCTTACTCTGTCNTCCTCTGTCTTTCTATCTCTCTGTAGCA	[G/C]

		AGAGG		
TN_301	GCAAGAGGAAAGAAGAAACATC A	GTCAGAG CCGAGAG TGGTCT	GCAAGAGGAAAGAAGAAACATCAACAA AGAAAGAGACATTGTCATGATAATGCCT GACTCAGTTTCACCCTGTNTGTGATCCTG AGACCACTCTCGGCTCTGAC	[A/G]
TN_2700	TAAACGGAAAG CCCAAGAAA	AGCGCGA GGTACTGT GTGTT	TAAACGGAAAGCCCAAGAAAATCAGAA AGCCAAGAACAATCTACTCCAGNTTTCA GCTCGCCGCCCTGCAGCGGAGATTTCAG AACACACAGTACCTCGCGCT	[T/C]
TN_1686	TGGGAAATAAG TAAACAAGTGT GG	CCAATGAA GTGATATG GACATTCT	TGGGAAATAAGTAAACAAGTGTGGGAT GGGCACATACTCAATACAGTGGTTTGCA TGAANGGAAAGATTTGTCCATATCTTGG ATAAGAATGTCCATATCACTTCATTGG	[A/G]
TN_2214	GTGGAGCGAGC CAACATT	TCCTTCAC CACCCTCT TCC	GTGGAGCGAGCCAACATTGGGCTGCCA CATTGTCGGCCCTCCAGGGTGAAGGAA GCGAGGGNGCGGAAGAGGGTGGTGAA GGA	[T/A]
TN_2076	CTGTGTCGTAG CAAGATGTGG	ACCTTGGC CTTTCTCA ACTG	CTGTGTCGTAGCAAGATGTGGCAGANG AAGTTGCCTTGGGGTTAAAATGCGAGAA GACTGTCCAGGGTCTTGCTCTATCTGCCA GCAGTTGAGAAAGGCCAAGGT	[T/G]
TN_2592	AGCATAGGATA GGTGAAAGCAA A	CCGTCCCT ACACATTG ATGA	AGCATAGGATAGGTGAAAGCAAATTCCT TGTGGACANCTACCTTTGGATAAACTGT AAAGTTTATCCTCATCAATGTGTAGGGA CGG	[C/G]
TN_666	TGATCCTGTGC AAATAAGAATG A	TGTGAATG GATATGA GGGACAC	TGATCCTGTGCAAATAAGAATGACTAGG ATTAGGACGCGAGGGGATGTTGTACAA GANGATATTAGGTCGAGAGTGTCCCTCA TATCCATTCACA	[T/G]
TN_744	GTCCTGCTGGT GTTGTGTTG	GCATCTTG GGTAGAG GATTTCA	GTCCTGCTGGTGTTGTGTTGCACCATAA CTTGACATAACCCTCAGAGTGNTAAACA GTAACTTGTGGGTCAAAGCCAGGACCTA TGGTGGAATTTGAAATCCTCTACCCAAG ATGC	[T/G]
TN_1980	AATTCAACAGC GAGCGAGTT	CGTCCGGC AAATCAGT AATC	AATTCAACAGCGAGCGAGTTTACNCTGA AACATTGCAAGGGAACCTATACCGTATT GGATTACTGATTTGCCGGACG	[A/G]
TN_2169	TCAAATGTTGTA GTTCTTCAGTTC G	AGACTGGT GTGGATA AGGAGAG A	TCAAATGTTGTAGTTCTTCAGTTCGACTG CCTCAGTTCTGGGACAACCCCTTTTTTTT CNGTCTCTCCTTATCCACACCAGTCT	[A/G]
TN_1597	AAAGGTTTGTC CCAGCCATA	ACAAGTG GCAGACG GGTTT	AAAGGTTTGTCCCAGCCATATTGTGCTA CCAGGCGTTGTCACCTGTGACTTGCGTT GCNGTCAAACCCGTCTGCCACTTGT	[T/C]
TN_192	ATTTGTTCCAG GCTTTGGTG	GCTGCTGT CTGTCTGT CCAA	ATTTGTTCCAGGCTTTGGTGGGGTTGAT GCGTATGAGGTGGGGATGATGGTANAA GAGCTGCAGATAGTGTTGGACAGACAG ACAGCAGC	[T/C]
TN_1350	ATTAACATTCTA ACCTTTGTCATT CG	GGTCCTGA GTTGATCT GTAGTGTA GT	ATTAACATTCTAACCTTTGTCATTCGAAA ANACCAATTCTTCATAGTACCAACTGATA GCATACTTTATGCTCCACCATTAAAAGTC AGTTAACTACACTACAGATCAACTCAGG ACC	[A/C]
TN_997	TCCAGCAGGTC ATAGATCAGAG	GCTCTTCG AGGTCGG ATG	TCCAGCAGGTCATAGATCAGAGAGCCCT TACCTCCGCCCAGNGCATCCAGGTCCTG ACATCCGACCTCGAAGAGC	[C/G]

TN_2181	GCCTTGATTTGATTAGCTCTGG	CAATGATGGACAGTTGTGGTG	GCCTTGATTTGATTAGCTCTGGTTTCTTGTACTTCTAATTACTCTCCACAAGAGGGTGGTCTTTAGCTATNAAGCGCACCACAACTGTCCATCATTG	[C/G]
TN_2462	GCTTCAGGATGTATCTGTGATGA	AGAGGGATTGGGATAAAGCTG	GCTTCAGGATGTATCTGTGATGAACAACTAAGCATTAACAATATCAATTNTATACAAGATCCAACCGAAGGTCCTTCAGCTTTATCCCAATCCCTCT	[T/G]
TN_1433	GACCCTGAGACCTGGATGG	AGCAGTGGCTTTCTGGTCAT	GACCCTGAGACCTGGATGGNTTTGGGCGCACTGTCTTTTGCTCTGGAGCTGGGAATCGGTGTGGATGACCAGAAAGCCACTGCT	[C/G]
TN_1016	GGGATTGGGAGAAACTGCTA	GGGATCTTTCCTGGGTCCT	GGGATTGGGAGAAACTGCTATTGGGAGAAACTGTTTTGTGTTAGGGAGACGACACTGNATGGGGAGAGCGGTGGTGGGGAGCCTACTGGAAAAGTAGGACCCAGGAAAGATCCC	[T/C]
TN_2423	TTGCACCTGGACACTGATCT	GGGTAGAAGTTGTCAGCCACA	TTGCACCTGGACACTGATCTATGGTCAGTTTTACATTTGCACCAATAAATGTTAAGGCTATGATGTGNGTGGGGTAGACTGATCCTAGGTCTGTGGCTGACAACTTCTACCC	[T/G]
TN_1110	AGACTTGTTGTCCTTTGACGTG	TGGAGGAGGTATTTATTGAAGTGA	AGACTTGTTGTCCTTTGACGTGTAGTTCTATCTGTTTGGATCCCAGATTACATTTTACTTCTTTGCCAGACGTCCCACGNGGTACATTCACTTCAATAAATACCTCCTCCA	[A/G]
TN_2074	TCCTCCTCGGCACTCTACTC	CGCACAGGTTCACAAAGGT	TCCTCCTCGGCACTCTACTCATGTAGACCGATAGCCGAGGCCTATATAAAGAATAGGCTTGATGTTTCCGACNAATGAAATGTAGCCTGGAATCTAACCTTTGTGAACCTGTGCG	[G/C]
TN_1822	AACATATACAGGAAGTGACGATTAACA	ACTGGCTGGAGAACCCTCA	AACATATACAGGAAGTGACGATTAACATCCTTAAGATAACGTGATGNGTGTTCCTGTGTGTCTCAGAACTCCTTGAGAGCGTCTCCTCCCTGAGGGTTCTCCAGCCAGT	[T/C]
TN_539	TTGACTTTGTGACCCTGCTG	CTGGTTCCTGTGCTGAAGGT	TTGACTTTGTGACCCTGCTGCAGCAACCAAAAATCATAGATACCTACTCTAGAATGGAGGAAGTATTCAACAACCTCACAGCACCCAGTGACCNCACCTTCAGCACAGGAACCAG	[A/G]
TN_2563	CTGCTGACCCTCTGACCTTC	GAATCTTGGCACAACACGTC	CTGCTGACCCTCTGACCTTCGACACTAGNGAAGCACCAGTGGCTCTACATCATCCCAGTACTTCTAAAGACGTGTTGTGCCAAGATTC	[A/G]
TN_2871	CCAACAACCAACAACCAACA	TGCGGACAATAGGGTGAGAT	CCAACAACCAACAACCAACAGACCTTGTCCTGTCCCCANCCTGCCAGTATGTCAGTAGATATAGTACCTCCGAAATCTCACCCTATTGTCCGCA	[T/C]
TN_2309	TTTCAGGTTGTCTTGCATCG	TCTCTGCGAATGTGCCATC	TTTCAGGTTGTCTTGCATCGACAGTTAAATGCTTAGGGCTCGATTCAATCTGTAGTGCTGAAGATCCACGCTACAGCGCAANATACATTTAAAGGCAATGATGGCACATTCGCAGAGA	[A/T]

APPENDIX H: The R script for genotyping functions.

```
pre.genotyping = function ( DIRname , CONTHR  , CORE , OUTseq  , OUTphred  , OUTcovs )  {

     # MODULE 1: Preliminary files/functions/libraries that are require to upload.
          {
               #libraries
               library (Biostrings) #This R package contains functions for doing modifications in teh DNA sequences.
               library(snow) # parallel computing

               #objects
               lowQ = strsplit(as.character(PhredQuality(0:18)),"")[[1]] # This is the object which defines the lowQ thershold level.

               revtag ="CATTAAGTTCCCATTA" # this is reverse tag at the end of reverse primer. This tag is targeted in PCR-2 as reverse site.
               fwdtag ="ACGACGTTGTAAAA" # this is forward tag at the end of reverse primer. This tag is targeted in PCR-2 as forward site. Any full length product seqeunced should have both seqeunce in the fastq file.

               # preparing barcode sequences to use in matching
               revbarc_i01to08 = c ( "CGTGATCATTAAGT" , "ACATCGCATTAAGT","GCCTAACATTAAGT","TGGTCACATTAAGT","CACTGTCATTAAGT","ATTGGCCATTAAGT","GATCTGCATTAAGT","TCAAGTCATTAAGT")
               revbarc_i01to08_comp = sapply(1:length(revbarc_i01to08), function(i) { as.character(reverseComplement( DNAString(revbarc_i01to08[i] ))) })
               revbarc_i01to08_names = paste("Ion-TrPi-i0",1:8,sep="")

               FILEname = list.files(path = DIRname, all.files = F,full.names = F)
               FILEname = FILEname [grepl ( "fastq$" , FILEname )]
               FILEnameINDEX = as.numeric(sapply ( FILEname , function(x) { substr ( x , 11, 13) } ))
          }
     # MODULE 1 ENDS.

     # MODULE 2: uploading FASTQ files.
          {
          listSeq2 = list()
          listPhred2 = list()

          for ( i in 1:length(FILEnameINDEX) ) { print(i)

               FastQ.1 = read.table( paste(DIRname,"/",FILEname[i],sep="") , stringsAsFactors=F , comment.char = "") # indiviual fastq files for each q. (we dont use th # charachter since it is included in the quality score as decsriptor)
               FastQ.seq = FastQ.1 [ seq(2,nrow(FastQ.1),by=4),1] # extract only only sequence from the fastq file. Note that one in every four lines are sequence,
```

```
one phred quality scores, and the other two are identifiers. We
dont need to have identifiers in te belwo code.
               FastQ.phred  = PhredQuality(FastQ.1 [
seq(4,nrow(FastQ.1),by=4),1]) # extract only only phred quality
score from the fastq file.

               identifierFWD = FILEnameINDEX[i]
               identifierREV = revbarc_i01to08_comp

               COVSindREV = sapply ( 1 : length(identifierREV) ,
simplify=F ,  function(j) {
                         ind1 = which ( vcountPattern
(identifierREV[j] , FastQ.seq , max.mismatch = 1, with.indels=T,
fixed=T) == 1 ) # this is match to reverse barcode #2
                         })
               COVSperREV = unlist(lapply ( COVSindREV , length
))
               REVS = which ( COVSperREV > CONTHR )

               listSeq2 [[ i ]] = list()
               listPhred2 [[ i ]] = list()

               for ( j in 1: length(REVS) ) {
                    listSeq2 [[ i ]] [[j]] = FastQ.seq [
COVSindREV [[ REVS[j] ]] ]
                    listPhred2 [[ i ]] [[j]] = FastQ.phred [
COVSindREV [[ REVS[j] ]] ]
               }

               names ( listSeq2 [[ i ]]  ) = REVS
               names ( listPhred2 [[ i ]]  ) = REVS

          }

          names(listSeq2) = FILEnameINDEX
          names(listPhred2) = FILEnameINDEX


     listSeq3 =  list()
     listPhred3 = list()
     k=1
     for ( i in 1 : length(listSeq2) ) { print(i)
          for ( j in 1 : length(listSeq2[[i]]) ) {

               listSeq3 [[k]] = listSeq2[[i]][[j]]
               names(listSeq3)[k] =
paste(names(listSeq2)[i],names(listSeq2[[i]])[j],sep = "_" )

               listPhred3 [[k]] = listPhred2[[i]][[j]]
               names(listPhred3)[k] =
paste(names(listPhred2)[i],names(listPhred2[[i]])[j],sep = "_" )

               k=k+1
          }}

     assign( eval(OUTseq) , listSeq3 )
     assign( eval(OUTphred) , listPhred3 )

     do.call ( save, list ( OUTseq , file= OUTseq ))
     do.call ( save, list ( OUTphred , file= OUTphred ))

     rm (listSeq2)
     rm (listPhred2)
```

```
     rm (listSeq3)
     rm (listPhred3)

     }
     # MODULE 2 ENDS.

     #~~~ MODULE 3: Calling nucleotide bases on the SNP position
~~~#
          {

          # This module is the major workhorse for calling the
genotypes. 1) Each targeted locus is called and quaility is
assesed.
          # This module uses raw sequence as input. "SET_FINAL2"
file is used to match primer, and SNP site sequences in the raw
seqeunces.
          # this function is slow (> 2-3 hours but less than a
day), so we use the R pacake "snow" which allows parallel
computing.
          cl <- makeSOCKcluster(rep("localhost",CORE)) # I have
eight cores in my work computer, and I allocated four-six of
them for this job.

          # objects, libraries used in parallel compueting
needs to be intruduced to each parallel nodes, by the below
function.

               load(file=OUTseq)
               load(file=OUTphred)

               clusterExport(cl, OUTseq , envir=environment() )
               clusterExport(cl, OUTphred , envir=environment()
)
               clusterExport(cl, "OUTseq", envir=environment() )
               clusterExport(cl, "OUTphred" ,
envir=environment())
               clusterExport(cl, "SET_FINAL2",
envir=environment())
               clusterExport(cl, "lowQ" , envir=environment() )
               clusterEvalQ(cl, library("Biostrings"))

          COVSpre = parSapply(cl , 1:length(get(eval(OUTseq))) ,
function(j) { print(j)      # function is run for each
individual (e.g. from 1 to length(get(eval(OUTseq))) )

               sapply(1:length(SET_FINAL2$FWD)  , function(i) {
print(i) # for each individual each loci run seperatlt.
Therfore, the belwo code is individual AND locus specific

                              SEQ1 = SET_FINAL2 [i,6]  # the
expected seqeunce including primers
                              FWD1 = SET_FINAL2 [i,4]  # forward
primer sequence
                              REV1 = SET_FINAL2 [i,7]  # reverse
primer seqeunce
                              Nind = SET_FINAL2 [i,12] # the
recognition seqeunce with SNP seqeunce marked as N.
                              reverseComplementIndex =
SET_FINAL2[i,10]==SET_FINAL2[i,11] # this is a compatibality
boolean for identify the SNP consistently with 7K illumina chip
calling. Ask Tutku for details.
                              Nind2 = Nind # back compatiablity
```

```
                            Nind3 = SET_FINAL2[i,13] # adapter
sequence. this and  te next line defines seqeunce to identfy tru
copy from the duplicates. In some cases the duplicate loci has
identical fwd and rev primers and recognition seqeunce, so we
take a sequence in the amplicon with fixed differences between
duplicates to only select the loci of interest.
                            Nind4 = SET_FINAL2[i,14] # anti
adapter

                       # the below function is to estimate
coverages. note tat ifelse is only for SDY (sex) loci.
                        covs2 = if (!is.na(Nind2)) {

                                  RC = vcountPattern ( REV1,
get(eval(OUTseq)) [[j]] , max.mismatch = 1, with.indels=T,
fixed=T ) # indices of seqeunces to reverse primer match (1
mismatch allowed).
                                  FC = vcountPattern ( FWD1,
get(eval(OUTphred)) [[j]] , max.mismatch = 1, with.indels=T,
fixed=T ) # indices of seqeunces to forward primer match (1
mismatch allowed).
                                     # note that one can do the
above routine by matching sequences to a truncated/shorter
primer seqeunce in the intesrtt of increasig coverage and in the
expense of specificity. Note that iontorrent specificity is low
esp. when polyN regions are present, therfore shorter
recognition seqeunces are better.

                                     CANDS =
get(eval(OUTseq))[[j]] [which( (RC+FC) > 0)]  # indices of
candidate sequences (at least one of rev or fwd primer matches
to a seqeunce)
                                     CANDSphred =
get(eval(OUTphred))[[j]] [which( (RC+FC) > 0)] # same as above
but in teh

                                       CANDS2 = CANDS
                                       CANDSphred2 = CANDSphred

                                     if (!is.na(Nind3)) { #
ifelse condition to accomodate "adapter" sequence
                                        CANDS2 = CANDS [grep
( Nind3, CANDS)]
                                        CANDSphred2 =
CANDSphred [grep ( Nind3, CANDS)]
                                        }

                                     if (!is.na(Nind4)) { #
ifelse condition to accomodate "antiadapter" sequence
                                        CANDS2 = CANDS
[!grepl ( Nind4, CANDS)]
                                        CANDSphred2 =
CANDSphred [!grepl ( Nind4, CANDS)]
                                        }

                                  # below four lines searches
for A,T,G,C in the SNP region.
                                     AAscanA = regexpr ( gsub
("N","A",Nind2) , CANDS2 )
                                     TTscanA =
regexpr(gsub("N",ifelse(reverseComplementIndex,"T","A"),Nind2),
CANDS2)
                                     GGscanA =
```

```
regexpr(gsub("N",ifelse(reverseComplementIndex,"G","C"),Nind2),
CANDS2)
                                        CCscanA =
regexpr(gsub("N",ifelse(reverseComplementIndex,"C","G"),Nind2),
CANDS2)

                                    # below four lines calculate
coverage for for each of the bases in the SNP region
                                          AAscanB = which (
AAscanA != -1)
                                        TTscanB = which ( TTscanA
!= -1)
                                        GGscanB = which ( GGscanA
!= -1)
                                        CCscanB = which ( CCscanA
!= -1)

                                    # below calculates coverage
in the "CANDS" other than the above four "specific" ones. In the
output it is represnted with a NUMBER. This is surrogate for
unspecific contribution by the primer.
                                        ALLscanB = length(CANDS)
     -length(c(AAscanB,TTscanB,GGscanB,CCscanB))

                                    # below marks each "specific"
coverage according to its quality. low quality (<20 phred score)
marks as 0,  others as 1.
                                        #MARKING A1
                                        AAs =
paste(as.numeric(!(substr(CANDSphred
[AAscanB],AAscanA[AAscanB],AAscanA[AAscanB])%in% lowQ)), rep("A"
, length(CANDS [AAscanB]) ),sep="_")
                                        #MARKING A2
                                        TTs =
paste(as.numeric(!(substr(CANDSphred
[TTscanB],TTscanA[TTscanB],TTscanA[TTscanB]) %in% lowQ)),
rep("T" , length(CANDS [TTscanB]) ),sep="_")
                                        #MARKING A1
                                        GGs =
paste(as.numeric(!(substr(CANDSphred
[GGscanB],GGscanA[GGscanB],GGscanA[GGscanB]) %in% lowQ)),
rep("G" , length(CANDS [GGscanB]) ),sep="_")
                                        #MARKING A2
                                        CCs =
paste(as.numeric(!(substr(CANDSphred
[CCscanB],CCscanA[CCscanB],CCscanA[CCscanB]) %in% lowQ)),
rep("C" , length(CANDS [CCscanB]) ) ,sep="_")

                                          # below is a fix to
evalute the quality ecore of teh base on teh SNP site, not the
SNP at the start of teh adapther seqeunce
                                          # noticed and suggested
by charlie waters
                                          #MARKING A1
                                          #AAs =
paste(as.numeric(!(substr(CANDSphred[AAscanB],AAscanA[AAscanB]+r
egexpr("N",Nind2),AAscanA[AAscanB]+regexpr("N",Nind2)) %in%
lowQ)), rep("A",length(CANDS[AAscanB])),sep="_")
                                          #MARKING A2
                                          #TTs =
paste(as.numeric(!(substr(CANDSphred[TTscanB],TTscanA[TTscanB]+r
egexpr("N",Nind2),TTscanA[TTscanB]+regexpr("N",Nind2)) %in%
lowQ)), rep("T",length(CANDS[TTscanB])),sep="_")
```

```
                                        #MARKING A1
                                        #GGs =
paste(as.numeric(!(substr(CANDSphred[GGscanB],GGscanA[GGscanB]+r
egexpr("N",Nind2),GGscanA[GGscanB]+regexpr("N",Nind2)) %in%
lowQ)), rep("G",length(CANDS[GGscanB])),sep="_")
                                        #MARKING A2
                                        #CCs =
paste(as.numeric(!(substr(CANDSphred[CCscanB],CCscanA[CCscanB]+r
egexpr("N",Nind2),CCscanA[CCscanB]+regexpr("N",Nind2)) %in%
lowQ)), rep("C",length(CANDS[CCscanB])),sep="_")


                                    ALLs =
c(AAs,TTs,GGs,CCs,ALLscanB)

                                #below else function is to
quantify the SDY coverage (sex loci). It doesnt have an "N" so
it is represented by a number in the output.

                                        } else {

                                        seqCOV  =
vcountPattern (SEQ1, get(eval(OUTseq))[[j]] , max.mismatch =
ifelse ( grepl("n",SEQ1), length(gregexpr("n",SEQ1,"")[[1]])+3,
3), with.indels=T, fixed=T)
                                        covs =
get(eval(OUTseq))[[j]] [which(seqCOV==1)]
                                        rep(1,length(covs))
     }

                            covs2
                })
          })

          colnames(COVSpre) = names(get(OUTseq))

          assign( eval(OUTcovs) , COVSpre )

          #do.call ( save, list ( OUTcovs , file=
paste(folder.to.create,"/",OUTcovs,sep="") ))
          do.call ( save, list ( OUTcovs , file= OUTcovs ))

          stopCluster(cl) # this one stops the parallel
computing

     }
     #~~~ MODULE 3 ENDS. ~~~#
 }
genotyping = function ( PRECOVFILE , lociBYlociPlot , THR ,
THRmvn , folder.to.create  ) {

     dir.create(folder.to.create)

     #~~~ Fast general descriptives ~~~#
     {

          readsVecR4 = unlist(PRECOVFILE) # vector form of the
coverage master file
          valid = sum (grepl("1",readsVecR4))
          invalid = sum (grepl("0",readsVecR4)) ##these fail the
quality
          PRECOVFILE[2,2]
          totCov_pre1 = length (readsVecR4)  #total coverage
```

```
real this number is taken from iontorrent server window (Heli1
and Heli2 runs)
          totCov_pre2 = c(na.exclude ( as.numeric ( readsVecR4 )
))
          totCov = totCov_pre1 + sum(totCov_pre2)

          valid/totCov # ~80 % valid ONTARGET
          invalid/totCov # 2 % invalid ONTARGET

     }
     #~~~ ENDS ~~~#

     #~~~ MODULE 3: COVERAGES ETC... ~~~#
     # This module outputs several objects. Some trivial while
others are more important for downstream anlayses or assay
descriptives/diagnostic.
     # Objects' row/column numbers are equal to number of
SNPs/individuals, respectively.
     {

          ## COVS1: Low quality bases, marked 0 (after phred
quality assesemnt) are removed from the dataset
      COVS1 = t(sapply( 1: nrow(PRECOVFILE), function(i) {
print(i)
           sapply(1: ncol(PRECOVFILE), function(j) {
                AA = PRECOVFILE[i,j][[1]]
                AA = AA [ is.na(as.numeric(AA))] #removing
NUMBER entry (number of seqeunces that contains either of the
primer but lacks the specific recognitions seqeunce).
                AA = if ( sum(is.na(AA))== length(AA) ) NA else
{
                     table(AA [!(is.na(AA) | grepl( 0, AA ))])
}
                AA = if( length(AA) ==0 ) NA else AA
                AA = if (is.numeric(PRECOVFILE[i,j][[1]]) )
sum(PRECOVFILE[i,j][[1]]) else AA
                list(AA)
                })
      }))

          ## COVS2 = Number of differnet type of nucleotide
bases (e.g. A,T,C, and G) called for specific SNP loci and
individual. NA= 0, One base type=1, Two base type= 2, Three base
type = 3 etc.. This is to filter out or understand the nature
SNP region. Third or forth bases may mean miscall,or duplicate
presense, among others)
      COVS2 = t(sapply( 1: nrow(PRECOVFILE), function(i) {
print(i)
           sapply(1: ncol(PRECOVFILE), function(j) {
                if ( sum(!is.na(COVS1[i,j][[1]]))==0 |
length(COVS1[i,j][[1]])==0 )     0 else {
length(COVS1[i,j][[1]]) }
      })
           }) )

     ## Loci taht has more than 2 allale called
     AA = which ( apply(COVS2,1,max) > 2) # locus with more
than expected genotype calls

          freqAL3 = sapply(AA, function(i) {
           AA2 = unlist(COVS1[i,])
           AA3 = AA2[grepl("1_",names(AA2))]
               AA4 = cbind(gsub("1_","",names(AA3)),AA3)
```

```
                AA5 = by(as.numeric(AA4[,2]),AA4[,1],sum)
                AA6 = sort(c(AA5),d=T)
                (1 - (sum(AA6[1:2]) / sum(AA6))) * 100
      })
      AA [ which(freqAL3>1) ] # All but two has thrid allel
genotype freqeny less than 1%
           SET_FINAL2 [ AA [ which(freqAL3>1) ]  , ] # TN_1746,
and TN_2606 has third allele, but this is from a paralogous loci
with a differnet SNP base. so exlduing it without hindering
resulst is easy.


           ## COV2b = Only expected nucleor=tides in the SNP site
are retained (using SET_FINAL2).
           ## vicky directed a fix here by adding ", simplify=F"
to the second supply function: If not implemented and when all
idividuals are heterozygote, the first spplly function genartes
a table, which then restrict the outer sapply tio become a
table, hence the code breaks...

      COVS2b = t(sapply( 1: nrow(PRECOVFILE),  function(i) {
print(i)
                      sapply(1: ncol(PRECOVFILE), simplify=F,
function(j) {
                            AA = COVS1[i,j][[1]]
                            AA = gsub("1_","",names(AA))
                            if(length(AA)==0) 0 else {
COVS1[i,j][[1]] [ AA %in% strsplit(SET_FINAL2[i,9],"")[[1]] ] }
                      })
      }))

      COVS2b[which ( SET_FINAL2[,2] == "SDY_ion2"),] = lapply (
PRECOVFILE[which ( SET_FINAL2[,2] == "SDY_ion2"),], as.numeric )
      ## adding SDY coverage maunally

      ## COVS3 = Coverages per ind/marker after incorrect calls
(NUMBERS)  removed.
      COVS3a = t(sapply( 1: nrow(PRECOVFILE), function(i) {
print(i)
            sapply(1: ncol(PRECOVFILE), function(j) {
                  sum(COVS1[i,j][[1]])
      })
            }) )

      ## COVS3b = Coverages per ind/marker after low quality
calls (0_N) AND incorrect calls (NUMBERS)  removed.
      COVS3b = t(sapply( 1: nrow(PRECOVFILE), function(i) {
print(i)
            sapply(1: ncol(PRECOVFILE), function(j) {
                  sum(COVS2b[i,j][[1]])
      })
            }) )

      # Below object is a boolean matrix to specify if (loci x
individuals) combination is assayed or not. In most cases all
individuals are genotyped in all SNPs, so all elements in the
matrix are 1.
           covMAT = matrix(1,nrow=nrow(PRECOVFILE),ncol=ncol
(PRECOVFILE))

           #COVS4 = IMPORTANT OBJECT. It shows coverage per
individual and locus, but it does not provide infromation on
genotype calls.
```

```
          #all loci are inculded for all samples
          COVS4a = COVS3a * covMAT
          COVS4b = COVS3b * covMAT
     COVS4b[which ( SET_FINAL2[,2] == "SDY_ion2"),]=
unlist(lapply (lapply ( PRECOVFILE[which ( SET_FINAL2[,2] ==
"SDY_ion2"),], as.numeric ),sum, na.rm=T )) # again adding SDY
coverage manually[which ( SET_FINAL2[,2] == "SDY_ion2"),]=
unlist(COVS1[which ( SET_FINAL2[,2] == "SDY_ion2"),]) # again
adding SDY coverage manually
          COVS4a[which ( SET_FINAL2[,2] == "SDY_ion2"),]=
unlist(lapply (lapply ( PRECOVFILE[which ( SET_FINAL2[,2] ==
"SDY_ion2"),], as.numeric ),sum, na.rm=T )) # again adding SDY
coverage manually[which ( SET_FINAL2[,2] == "SDY_ion2"),]=
unlist(COVS1[which ( SET_FINAL2[,2] == "SDY_ion2"),]) # again
adding SDY coverage manually

          sum(COVS4b,na.rm=T) / totCov  # on target coverage

     # again, some descriptives

           sum(COVS4b)

           sum(COVS4b[-6,]>10,na.rm=T) / prod(dim(COVS4b[-6,]))
# 93 % is higher than 10x coverage
           sum(COVS4b[-6,]>13,na.rm=T) / prod(dim(COVS4b[-6,]))
# 90 % is higher than 13x coverage
           sum(COVS4b[-6,]>30,na.rm=T) / prod(dim(COVS4b[-6,]))
# 65 % is higher than 20x coverage

               #coverege per indvidual
               aveCOV1 = round (sapply ( 1:dim(COVS4b)[2],
function(i) { sum(unlist(COVS1[,i]),na.rm=T) / dim(COVS4b)[1] })
)
               #coverege per locus
               aveCOV2 = round (sapply ( 1:dim(COVS4b)[1],
function(i) { sum(unlist(COVS1[i,]),na.rm=T) / dim(COVS4b)[2] })
)


          }
     #~~~ MODULE 3 ENDS ~~~#

     #~~~ MODULE GENOTYPING ~~~#
     {
          # Here we adopted campbell et al 2015 method to call
for genotypes. See their figure 4 (or 5) for more details.
          # We adopted a "transformation of axis" based
methodllogy to account for MVN loci with stringent criteria for
calling genotypes.

          COVSg1 = COVS2b

          #this matrix counts valid As only.
          COVSg1_A = t(sapply( 1: nrow(COVSg1), function(i) {
print(i)
                         sapply( 1: ncol(COVSg1), function(j)  {
                              A =
as.numeric(COVSg1[i,j][[1]]["1_A"] )
                              if (is.na(A)) 0 else A
                              })   }))
          #this matrix counts valid Ts only.
          COVSg1_T = t(sapply( 1: nrow(COVSg1), function(i) {
print(i)
```

```
                        sapply( 1: ncol(COVSg1), function(j)  {
                            A =
as.numeric(COVSg1[i,j][[1]]["1_T"] )
                            if (is.na(A)) 0 else A
                            })   }))
          #this matrix counts valid Gs only.
          COVSg1_G = t(sapply( 1: nrow(COVSg1), function(i) {
print(i)
                        sapply( 1: ncol(COVSg1), function(j)  {
                            A =
as.numeric(COVSg1[i,j][[1]]["1_G"] )
                            if (is.na(A)) 0 else A
                            })   }))
          #this matrix counts valid Cs only.
          COVSg1_C = t(sapply( 1: nrow(COVSg1), function(i) {
print(i)
                        sapply( 1: ncol(COVSg1), function(j)  {
                            A =
as.numeric(COVSg1[i,j][[1]]["1_C"] )
                            if (is.na(A)) 0 else A
                            })   }))

          COVSg1_A = COVSg1_A * covMAT
          COVSg1_T = COVSg1_T * covMAT
          COVSg1_G = COVSg1_G * covMAT
          COVSg1_C = COVSg1_C * covMAT

          AA = apply(COVSg1_A,1, function(x) { mean(x,na.rm=T) }
)
          TT = apply(COVSg1_T,1, function(x) { mean(x,na.rm=T) }
)
          GG = apply(COVSg1_G,1, function(x) { mean(x,na.rm=T) }
)
          CC = apply(COVSg1_C,1, function(x) { mean(x,na.rm=T) }
)

          #average coverage per loci, per base.
          round(cbind(AA,TT,GG,CC))

          ### this to avoid infinite numbers when estimating
coverages. (based on campbell et al 2004)
          COVSg1_A [COVSg1_A==0] = 0.1
          COVSg1_T [COVSg1_T==0] = 0.1
          COVSg1_G [COVSg1_G==0] = 0.1
          COVSg1_C [COVSg1_C==0] = 0.1

          GENCALL1cov =  matrix(NA, nrow(COVSg1),ncol(COVSg1) )
          GENCALL2cov =  matrix(NA, nrow(COVSg1),ncol(COVSg1) )
          GENCALL =  matrix(NA, nrow(COVSg1),ncol(COVSg1) )
          GENCALLnumbers =  matrix(NA, nrow(COVSg1),ncol(COVSg1)
)

          # Transfromation of MVN is given in SET_FINAL2$MVNlike
column. If 0 it is not MVN, other values gives tranfromation
angle

          # this double loop fills in coverage of specific SNPs
per individuals
          # current thrshold setting normal loci: coverege = 10x
, pp < 0.1, qq > 10, 0.2 < pq < 5
          # current thrshold setting MVN loci: coverege = 20x ,
pp < 0.67, qq > 12.5, 0.25 < pq < 4 (much stringent calling
criteria for MVN)
```

```
          # -9 depits fail due to not passing the therhold. NA
depits failiure due to failure the proprtion of allele coverage
thrhold


     for ( i in 1:nrow(COVSg1) ) { print(i) # this one runs
locus by locus

          for ( j in 1:ncol(COVSg1) ) {  # this one runs
individuals by individuals. SO each locus (index "i") is
executed for every individual (insex "j")

               if (SET_FINAL2$MVNlike[i] == 0) {

                              alleles = strsplit ( SET_FINAL2[i,9]
, "" )[[1]] # selects only targeted bases at the selected loci

                                VAR = which(c("A","T","G","C")
%in% alleles) # targeted alleles are selected here

                                A1 = c("A","T","G","C") [VAR][1] #
tareget allele 1
                              A2 = c("A","T","G","C") [VAR][2] #
tareget allele 2

                              COVSggg = c(COVSg1_A [i,j],COVSg1_T
[i,j],COVSg1_G [i,j],COVSg1_C [i,j])
                              COVSggg2 = COVSggg [VAR] #only
infromation related to targetd alleles are kept


                                #alleles = strsplit (
SET_FINAL2[i,9] , "" ) [[1]] # selects only targeted bases at
the selected loci
                              #alleles2 = alleles [alleles %in%
LETTERS]

                                #A1 = alleles2[1] # tareget allele
1
                              #A2 = alleles2[2] # tareget allele 2

                              #COVSggg =  c(COVSg1_A
[i,j],COVSg1_T [i,j],COVSg1_G [i,j],COVSg1_C [i,j])
                              #COVSggg2 = COVSggg
[c("A","T","G","C") %in% alleles2] #only infromation related to
targetd alleles are kept

                              VALgg = COVSggg2[1] / COVSggg2[2]

                              VALgg2 = if ( is.na(VALgg) ) NA else
{
                                                  if(  (COVSggg2[1] +
COVSggg2[2])  < THR ) -9 else { ## loci with less than 10
coverage is not called
                                                        if( VALgg<0.1 )
paste(A2,A2,sep="") else { ## ratio of coverage <0.1 and >10 are
called homozygote for each allele
                                                              if(
VALgg>10 ) paste(A1,A1,sep="") else {
                                                                    if(
VALgg>0.2 & VALgg<5) paste(sort(c(A1,A2)),collapse="",sep="")
else {NA} }}}} ## coverege proprtion between 0.2. and 5 are
called heterzygote.
```

```
                            VALgg2numbers = if ( is.na(VALgg) )
NA else {
                                                if(  (COVSggg2[1] +
COVSggg2[2])  < THR ) -9 else {
                                                      if( VALgg<0.1 )
0 else {
                                                            if(
VALgg>10 ) 2 else {
                                                                  if(
VALgg>0.2 & VALgg<5) 1 else {NA} }}}}

                          GENCALL1cov [i,j] = COVSggg2[1]
                       GENCALL2cov [i,j] = COVSggg2[2]
                       GENCALL [i,j] = VALgg2
                       GENCALLnumbers [i,j] = VALgg2numbers

                     }

                  if (SET_FINAL2$MVNlike[i] > 0)  {

                   alleles = strsplit ( SET_FINAL2[i,9] , ""
)[[1]] # selects only targeted bases at the selected loci

                      VAR = which(c("A","T","G","C") %in% alleles)
# targeted alleles are selected here

                      A1 = c("A","T","G","C") [VAR][1] # tareget
allele 1
                   A2 = c("A","T","G","C") [VAR][2] # tareget
allele 2

                   COVSggg =  c(COVSg1_A [i,j],COVSg1_T
[i,j],COVSg1_G [i,j],COVSg1_C [i,j])
                   COVSggg2 = COVSggg [VAR] #only infromation
related to targetd alleles are kept

                   degree1 = ( 180 / ( pi / atan ( COVSggg2[2] /
COVSggg2[1] )  ))
                      degree2 = round ( ifelse ( degree1 > 90, 90,
ifelse ( degree1 < SET_FINAL2$MVNlike[i] , SET_FINAL2$MVNlike[i]
, degree1 )))
                      KK = round ( seq(SET_FINAL2$MVNlike[i]
,90,by=1) * seq(1,0,length.out =
length(SET_FINAL2$MVNlike[i]:90)) )
                      seq(SET_FINAL2$MVNlike[i] ,90,by=1)

                      KK2 = KK [ which ( degree2 ==
seq(SET_FINAL2$MVNlike[i] ,90,by=1) )]

                      degree3 = degree2-KK2
                      R = sqrt ( (COVSggg2[2])^2 + (COVSggg2[1])^2
)
                      YYY1 = round ( sin ( pi * (degree3/180) ) *
R )
                      XXX1 = round ( cos ( pi * (degree3/180) ) *
R )
                      c(XXX1,YYY1)

                      COVSggg2[1] = XXX1
                   COVSggg2[2] = YYY1

                            VALgg = COVSggg2[1] / COVSggg2[2]
```

```
                            VALgg2 = if ( is.na(VALgg) ) NA else
{
                                                    if(  (COVSggg2[1] +
COVSggg2[2])  < THRmvn ) -9 else { ## loci with less than 20
coverage is not called
                                                        if( VALgg<0.067
) paste(A2,A2,sep="") else { ## ratio of coverage <0.067 and
>12.5 are called homozygote for each allele
                                                            if(
VALgg>12.5 ) paste(A1,A1,sep="") else {
                                                                if(
VALgg>0.25 & VALgg<4) paste(sort(c(A1,A2)),collapse="",sep="")
else {NA} }}}} ## coverege proprtion between 0.25. and 4 are
called heterzygote.

                            VALgg2numbers = if ( is.na(VALgg) )
NA else {
                                                    if(  (COVSggg2[1] +
COVSggg2[2])  < THRmvn ) -9 else {
                                                        if( VALgg<0.067
) 0 else {
                                                            if(
VALgg>12.5 ) 2 else {
                                                                if(
VALgg>0.25 & VALgg<4) 1 else {NA} }}}}

           GENCALL1cov [i,j] = COVSggg2[1]
           GENCALL2cov [i,j] = COVSggg2[2]
           GENCALL [i,j] = VALgg2
           GENCALLnumbers [i,j] = VALgg2numbers


               }

               if (SET_FINAL2$MVNlike[i] < 0)  {

                   alleles = strsplit ( SET_FINAL2[i,9] , ""
)[[1]] # selects only targeted bases at the selected loci

                   VAR = which(c("A","T","G","C") %in% alleles)
# targeted alleles are selected here

                   A1 = c("A","T","G","C") [VAR][1] # tareget
allele 1
                A2 = c("A","T","G","C") [VAR][2] # tareget
allele 2

                COVSggg =  c(COVSg1_A [i,j],COVSg1_T
[i,j],COVSg1_G [i,j],COVSg1_C [i,j])
                COVSggg2 = COVSggg [VAR] #only infromation
related to targetd alleles are kept
                COVSggg2 =  rev(COVSggg2)

                degree1 = 180 / ( pi / (atan ( COVSggg2[2] /
COVSggg2[1] )  ))
                   degree2 = round ( ifelse ( degree1 > 90, 90,
ifelse ( degree1 < abs(SET_FINAL2$MVNlike[i]) ,
abs(SET_FINAL2$MVNlike[i]) , degree1 )))

                   KK = (round ( seq(
abs(SET_FINAL2$MVNlike[i]) ,90,by=1) * seq(1,0,length.out =
length( abs(SET_FINAL2$MVNlike[i]) :90)) ))
```

```
                    KK2 = KK [ which ( degree2 == seq(
abs(SET_FINAL2$MVNlike[i]) ,90,by=1) )]

                    degree3 = degree2-KK2
                    R = sqrt ( (COVSggg2[2])^2 + (COVSggg2[1])^2
)
                    YYY1 = round ( sin ( pi * (degree3/180) ) *
R )
                    XXX1 = round ( cos ( pi * (degree3/180) ) *
R )
                    c(XXX1,YYY1)

                    COVSggg2[1] = YYY1
                COVSggg2[2] = XXX1
                          VALgg = COVSggg2[1] / COVSggg2[2]

                          VALgg2 = if ( is.na(VALgg) ) NA else
{
                                              if(  (COVSggg2[1] +
COVSggg2[2])  < THRmvn ) -9 else { ## loci with less than 20
coverage is not called
                                                  if( VALgg<0.067
) paste(A2,A2,sep="") else { ## ratio of coverage <0.067 and
>12.5 are called homozygote for each allele
                                                      if(
VALgg>12.5 ) paste(A1,A1,sep="") else {
                                                          if(
VALgg>0.25 & VALgg<4) paste(sort(c(A1,A2)),collapse="",sep="")
else {NA} }}}} ## coverege proprtion between 0.25. and 4 are
called heterzygote.

                          VALgg2numbers = if ( is.na(VALgg) )
NA else {
                                              if(  (COVSggg2[1] +
COVSggg2[2])  < THRmvn ) -9 else {
                                                  if( VALgg<0.067
) 0 else {
                                                      if(
VALgg>12.5 ) 2 else {
                                                          if(
VALgg>0.25 & VALgg<4) 1 else {NA} }}}}

           GENCALL1cov [i,j] = COVSggg2[1]
           GENCALL2cov [i,j] = COVSggg2[2]
           GENCALL [i,j] = VALgg2
           GENCALLnumbers [i,j] = VALgg2numbers


                    }

               }
     }

          dim(GENCALL1cov)
          GENCALL1cov[1:10,1:10] # coverage allele 1
          GENCALL2cov[1:10,1:10] # coverage allele 2
          GENCALL[1:10,1:10] # genotype call (-9 for no call)
          GENCALLnumbers[1:10,1:10] # number of diffrenet bases
in the call

          #(snmall descriptive) call failiure is larger in MVN
due to stringent criteria
          normalloci = as.vector(GENCALL
```

```
[which(SET_FINAL2$MVNlike==0),])
          sum(is.na(normalloci)) / length(normalloci)
          MVNloci = as.vector(GENCALL
[which(SET_FINAL2$MVNlike!=0),])
          sum(is.na(MVNloci)) / length(MVNloci)

          #genotype success rate per individuals
          genoSuccessInd = sapply( (1:ncol(GENCALL)),
function(j) {
               AA1 = GENCALL[,j]
           AA2 = AA1 [!is.na(AA1)]
           AA3 = 1-sum(grepl("-9",AA2))/length(AA2)
          })

          #genotype success rate per loci

          genoSuccessLoci = sapply( (1:nrow(GENCALL))[ ] ,
function(i) { #1:nrow(GENCALL))[-8]
           AA1 = GENCALL[i,] # GENCALL[i,-8]
           AA2 = AA1 [!is.na(AA1)]
           AA3 = 1-sum(grepl("-9",AA2))/length(AA2)
          })

          genoSuccessLoci [which ( SET_FINAL2[,2] ==
"SDY_ion2")] = 1 ## sdy sexing loci is eqauled to 1
     }
     #~~~ MODULE GENOTYPING ENDS~~~#

     #~~~ MODULE SEXING ~~~#
     {
          # males tat has low genotying sucess may also have low
to zero coverage. Therefore they should be excldued.
          # we excldued individuals that has lower than 80%
genotyoing sucess to exclude incorrect calling due to low
genotyoing sucess. (therodhold can be change later based on
expereience)
          # we can set two tehshold, such as 0.1 and 0.3 and
leave in between unassigned. These therhoplds are arbitararyly
selected.

          aveCOV1 = round (sapply ( 1:dim(COVS4b)[2],
function(i) { sum(unlist(COVS1[,i]),na.rm=T) / dim(COVS4b)[1] })
)
          aveSDY = (unlist ( COVS4b [ SET_FINAL2[,2] ==
"SDY_ion2"  , ] ) / aveCOV1) # individual SDY coverage normlized
to mean individual coverage
          highGenoSucInd = which(genoSuccessInd>0.70) # only
individuals with genoSuccessInd>0.90 is selected.
          #plot(aveSDY[highGenoSucInd],pch=20,ylim=c(0,4)) #
CLEAR BIMODALITY OBSERVED!!
          #abline(h=0.80)
          #abline(h=0.15)
          genoSEX_PS = rep(NA, length(aveSDY) )
          #genoSEX_PS [aveSDY>0.40]=1
          #genoSEX_PS [aveSDY<0.15]=0
          #table(genoSEX_PS,useNA="a")
          genoSEX_PS [ highGenoSucInd ] =  round( aveSDY [
highGenoSucInd ] , 3)
          #plot ( GENCALL[6,] , ylim=c(0,3))
          GENCALLnumbers[6,]= genoSEX_PS
          GENCALL[6,]= genoSEX_PS
     }
     #~~~ MODULE SEXING ENDS~~~#
```

```
     ## A) FIGURE TO INVESTIGAE PER LOCUS TOPOGRPGY OF COVERAGE
     {
          ## below objects are from "Aykanat et al 2016 JFB main
code 040916" line 477-478.
          ## it contains info on coverage numbers of each of teh
alternative allele.
          ## note that this is after correcting for MVN
structure etc..
          ## please use "COVS1" object in "Aykanat et al 2016
JFB main code 040916" at line 252 and modify code for raw
calculations .
          ## note taht coverage info is ignored for sdy loci,
which had no alternative allale.

           GENCALLnumbers[1:10,1:10] ## this is for colors
          MVNlike = ifelse(SET_FINAL2$MVNlike==0, "no MVN",
"MVN" )

          #simple image example (of 10th loci which is
SET_FINAL[10])
          #plot(GENCALL1cov[10,], GENCALL2cov[10,], pch=20)
          #abline(0,1) # this is expected heterzygot line

     if (   lociBYlociPlot == T ) {
          # plotting all allales in a pdf file
          pdf(file = paste(folder.to.create,
"/lociBYlociCOV%03d.pdf",sep="") , , paper= "a4", width = 7,
height = 11)
          par(mfrow=c(7,4),mar=c(2,2,1,1))
          for (i in (1: 197 ) ) { print(i)

               MX = max(c(GENCALL1cov[i,],GENCALL1cov[i,]))

                    COL1 = GENCALLnumbers[i,]
                    COL1 [which(COL1==-9)] = "black" # this
uncalled for failing criteria to call between genotypes
                    COL1 [is.na(COL1)] = "dark gray"  # this
uncalled due to low coverage
                    COL1 [which(COL1==0)]="blue"
                    COL1 [which(COL1==1)]="purple"
                    COL1 [which(COL1==2)]="red"

               if(i == 6 ) { plot(1,1, main = "this is SDY") }
else {

               plot( GENCALL1cov[i,] , GENCALL2cov [i,]
,col=COL1 ,  pch= 1, xlim=c(0,MX), ylim=c(0,MX) ,
                    cex=0.6, cex.axis=0.7 , main = paste(i,",",
SET_FINAL2[i,2],",", SET_FINAL2[i,3],",", MVNlike [i]), cex.main
= 0.6, ylab=NA,xlab=NA)
                    # if you re-adjust thersholds in teh main
code when calling genotypes, please adjust them accordingly here
as well.
                    if (MVNlike [i] != "no MVN")  { abline(0
,0.1 , lty=3,col="blue");abline(0,0.2,lty=3,col="purple"
);abline(0,5,lty=3,col="purple");abline(0,10,lty=3, col="blue"
);abline(THRmvn,-1,lty=1,lwd=2);abline(0,1,lwd=2,col="purple") }
else {abline(0,0.067
,lty=3,col="blue");abline(0,0.25,lty=3,col="purple");abline(0,4,
lty=3,col="purple");abline(0,12.5,lty=3,col="blue");abline(THR,-
1,lty=1,lwd=2);abline(0,1,lwd=2,col="purple") }
               }
```

```
                }
          dev.off()
          ## colors indicate called genotyped in "Aykanat et al
2016 JFB main code 040916". inspect the figure by eye if you are
genotyping a diffrenet population, or if you are adding new
pimers here.
     }
     }
     ## A) ENDS

     pdf(file = paste(folder.to.create, "/coverage by loci and
individual.pdf" , sep="") , paper= "a4", width = 7, height = 11)
     par(mfrow=c(2,1))
     plot(log(sort(aveCOV2)),xlab="loci sorted by
cov",ylab="coverage",pch=20,yaxt="n")
     axis (2 , at  =  log ( c(10,20,50,100,250) ) , labels =
c(10,20,50,100,250) , las=2)
     plot(log(sort(aveCOV1)),xlab="individuals sorted by
cov",ylab="coverage",pch=20,yaxt="n")
     axis (2 , at  =  log ( c(10,20,50,100,250) ) , labels =
c(10,20,50,100,250) , las=2 )
     dev.off()

     pdf(file = paste(folder.to.create, "/genotyping success by
loci and individual.pdf",sep="") , paper= "a4", width = 7,
height = 11)
     par(mfrow=c(2,1))
     plot ( sort ( genoSuccessInd ) , ylab = "individual
genotyping success",cex=0.1,pch=20)
     abline( h=0.9)
     plot ( sort ( genoSuccessLoci ), ylab = "locus genotyping
success",cex=0.1,pch=20)
     abline( h=0.9)
     dev.off()

     average.coverage.by.loci = aveCOV2
     average.coverage.by.individuals = aveCOV1
     genotyping.success.by.individual = genoSuccessInd
     genotyping.success.by.loci = genoSuccessLoci

     genotyping.byloci =
cbind(SET_FINAL2$id,SET_FINAL2$name,SET_FINAL2$SNP,average.cover
age.by.loci,genotyping.success.by.loci)
     genotyping.individual = cbind(colnames(PRECOVFILE),
average.coverage.by.individuals,genotyping.success.by.individual
)

     summary.geno = list ( GENCALL1cov , GENCALL2cov ,
GENCALLnumbers , GENCALL ,  genotyping.byloci ,
genotyping.individual)

     for(i in 1:4) {
          colnames(summary.geno[[i]]) = colnames(PRECOVFILE)
          rownames(summary.geno[[i]]) =   SET_FINAL2$id    }


     save(summary.geno ,  file=
paste(folder.to.create,"/summary.geno",sep="") )
     write.table ( summary.geno [[3]] , file =
paste(folder.to.create,"/genotype.by.numbers.txt",sep=""),
sep="\t",quote = F)
     write.table ( summary.geno [[4]] , file =
paste(folder.to.create,"/genotype.by.bases.txt",sep=""),
```

```
sep="\t",quote = F)
     write.table ( summary.geno [[5]] , file =
paste(folder.to.create,"/genotyping.descriptives.byloci.txt",sep
=""), sep="\t",quote = F)
     write.table ( summary.geno [[6]] , file =
paste(folder.to.create,"/genotyping.descriptives.individual.txt"
,sep=""), sep="\t",quote = F)
     README = " We have four outputs here: 1)
'genotype.by.numbers.txt', 2) 'genotype.by.bases.txt', 3)
'genotyping.descriptives.byloci.txt', 4)
'genotyping.descriptives.individual.txt', 5) 'coverage by loci
and individual.pdf',  'genotyping success by loci and
individual.pdf' and (optional) 'lociBYlociCOV%03d.pdf' figures,
6) An R object (list) called 'summary.geno', which contains all
'txt' output and total coverage per allele per individuals."
     write.table(README, file = paste(folder.to.create,
"/README",sep=""),sep = "\t", quote = F)
```

APPENDIX I: The R script used to run the genotyping functions (Appendix H), merge the data from the re-run of the 6th library and include the additional 46 adult Atlantic salmon to the final dataset (object labeled as GG_JL_All_Gy).

```
load(file="SET_FINAL")   # this file is in the wroking
directory. This is an important tabular object which provides
primers, product sequence, SNP variant, and other relevant
infromation of loci in the panels. This object is used to match
raw reads to target loci and to call the SNP variant
(genotyping). A target locus is matched using forward and
reverse primer infromation, as well as the recognition sequence,
which is the seqeunce information around the SNP site. Note
that, for a few SNPs, the way the recognition sequence selected
deviates from from the description in the MS, and for very few
cases an  adapter seqeunces has been included to further improve
genotype topogrphy (e.g. figure 5 in the paper). Finally, the
total number of SNPs in this object is 197. About 20 SNPs from
Aykanat et al 2016 (10% of 216 SNP) is not included here. Our
post-acceptance results using a larger number of individuals
suggests genotype calls of some SNPs exhibited some
disconcordance between 7K data and ion-torrent geneotypigng.
Therfore, we do not advice to include thsose SNPs in the
anlaysis. SNPs provided here had >99% conconrdance with 7K SNP
data across 192 individuals (unpublished data). Likeiwise, SNPs
taken from 220K SNPS (i.e. sea age SNPs) has high (>99%)
conconrdance. Finally, this diagnostoc test is optimised for
Teno river Atlantc salmon and slight changes in genotype calling
topography are likley when used with phenogentcially distant
lineages. We advice an intial visual inspection of coverage
topogrpagy.
SET_FINAL2 = SET_FINAL
save (SET_FINAL2 , file = "SET_FINAL2")

##
folders = list.files() [1:9]

## loading raw outout and pre-processing
for( RR in 1: length(folders) ) { print(RR)
source ("genotyping.functions.250717.R")  # this file is in the
wroking directory
#INPUT FOR "pre.genotyping" function. ALL NEEDS TO BE DEFINED
     DIRname = folders[RR] # the folder that contains fastq
files. It must be located under the working directory OR FULL
PATH NEEDS TO BE GIVEN.
     #OR DIRname =
setwd("C:/Users/tutayk/Desktop/BALSA/TSP_SAMPLE_OPT_BARCODE")
     CONTHR = 1000 ## the coverage for reverse barcode
considered as valid (keep it at default).
     CORE = 4 ## number of cores to allocatae for genotyping
     OUTseq = paste ( gsub("user
","",folders[RR]),"listSeqTRIAL",sep="_") ## specify the name of
the list that will store sequences by indviudal ID.
     OUTphred = paste ( gsub("user ","",folders[RR])
,"listPhredTRIAL",sep="_") ## specify the name of the list that
will store sequence quality scores by indviudal ID.
     OUTcovs = paste ( gsub("user ","",folders[RR])
,"COVSpreTRIAL",sep="_") ## specify the name of the list that
will store pre-coverage files. This is teh file taht will be
used by "genotyping" function.

     pre.genotyping ( DIRname  , CONTHR , CORE , OUTseq  ,
OUTphred , OUTcovs )
```

```
     # OUTPUT are "R objects" named "listSeqTRIAL",
"listPhredTRIAL" and "COVSpreTRIAL".
     # "COVSpreTRIAL" is input for "genotyping" function.
}

### scoring genotypes ##
     for( RR in 1: length(folders) ) { print(RR)
          setwd("C:/Users/tutayk/Desktop/JanLaine")
          source ("genotyping.functions.250717.R")  # this file
is in the wroking directory

          load(file=
paste("C:/Users/tutayk/Desktop/JanLaine/",gsub ( "user " , "" ,
folders[RR]),"_COVSpreTRIAL",sep=""))
          PRECOVFILE = get(paste ( gsub ( "user " , "" ,
folders[RR]),"_COVSpreTRIAL",sep="" ))
          lociBYlociPlot = T # a boolean by which you decide to
have these plots or not.
          THR = 10 # the thrshold which a genotype will not be
called for an individuals if coverage is less than this.
          THRmvn = 15 # same as above, but defined for MVN plot.
Note that we keep mvn tehrhold a bit higher.
          folder.to.create = paste( folders[RR],
"/genotyping.resultsANDfigures", sep="") # a folder will be
cretaed by "genotyping" function, in which results will be
stored.
          genotyping ( PRECOVFILE, lociBYlociPlot, THR , THRmvn,
folder.to.create)
     }
### ENDS: scoring genotypes##
### compiling genotype scores of individuals ##
{
     KK=1
     GG_JL_All = sapply(1: length(folders), simplify=F ,
function(KK) {
          GG = read.table ( stringsAsFactors = F, sep = "\t" ,
h=T ,
          paste ( "C:/Users/tutayk/Desktop/JanLaine/",
folders[KK] ,
"/genotyping.resultsANDfigures/genotype.by.numbers.txt" , sep =
"")
          )

          colnames(GG) = gsub ( "X" , gsub ( "user " , "" ,
paste ( folders[KK], "_" , sep="") ) , colnames(GG) )
          t(GG)
     })
          summary ( GG_JL_All )
          dim ( GG_JL_All [[1]])
     # combining runs to a single data frame
          GG_JL_All_B  = do.call(  rbind , GG_JL_All)
          dim(GG_JL_All_B)
          head(GG_JL_All_B)

                GG_JL_All_BASE = sapply(1: length(folders),
simplify=F , function(KK) {
                     GG = read.table ( stringsAsFactors = F, sep
= "\t" , h=T ,
                     paste ( "C:/Users/tutayk/Desktop/JanLaine/",
folders[KK] ,
"/genotyping.resultsANDfigures/genotype.by.bases.txt" , sep =
"")
```

```
                    )

                    colnames(GG) = gsub ( "X" , gsub ( "user " ,
"" , paste ( folders[KK], "_" , sep="") ) , colnames(GG) )
                    t(GG)
               })
                    summary ( GG_JL_All_BASE )
                    dim ( GG_JL_All_BASE [[1]])
               # combining runs to a single data frame
                    GG_JL_All_BASE_B  = do.call(  rbind ,
GG_JL_All_BASE)
                    dim(GG_JL_All_BASE_B)
                    head(GG_JL_All_BASE_B)

     ## IDrows by run name, fwd and rev barcodes
     rownames(GG_JL_All_B) = gsub ("2nd_A","2ndA",
rownames(GG_JL_All_B) )
     rownames(GG_JL_All_B) = gsub ("2nd_B","2ndB",
rownames(GG_JL_All_B) )

     rownames(GG_JL_All_BASE_B) = gsub ("2nd_A","2ndA",
rownames(GG_JL_All_BASE_B) )
     rownames(GG_JL_All_BASE_B) = gsub ("2nd_B","2ndB",
rownames(GG_JL_All_BASE_B) )

     IDrows =  ( do.call ( rbind , strsplit (
rownames(GG_JL_All_B) , split = "_" ) ) )
     ##

     table(GG_JL_All_B[,-6],useNA ="a")
     FF = names ( which (  ( apply ( is.na(GG_JL_All_B[,-6]) , 2
, sum , na.rm=T ) / 2268) > 0.01) )
     FF

     ## combning IDrows and genotyep scores
     GG_JL_All_C = cbind (IDrows , GG_JL_All_B )
     GG_JL_All_BASE_C = cbind (IDrows, GG_JL_All_BASE_B)
     head (GG_JL_All_C)
     dim (GG_JL_All_C)
     #
     table (GG_JL_All_C [ , c (4:200)[-6] ] , useNA = "a")
     # -9 scores to NA
     GG_JL_All_C [GG_JL_All_C==-9 ]=NA
     GG_JL_All_BASE_C [GG_JL_All_BASE_C==-9 ]=NA
}
### ENDS: genotype scores of individuals ##

GG_JL_All_C ## FILE ALL
dim(GG_JL_All_C)
dim(GG_JL_All_BASE_C)

GG_JL_All_C[1:10,1:5]
table(GG_JL_All_C[,1])

length(rownames(GG_JL_All_C))
length(unique(rownames(GG_JL_All_C)))

GG_compositeID = paste ( GG_JL_All_C[,1] , GG_JL_All_C[,2] ,
GG_JL_All_C[,3] , sep = "_" )

### infromation file now ##

infoJL = read.table ( stringsAsFactors = F , h =T , sep = "\t" ,
```

```
file = "C:/Users/tutayk/Desktop/JanLaine/IT-runs sample details
for Tutku.txt")
dim(infoJL)
head(infoJL)

infoJL [ which ( infoJL$run == "3rd_1MP" ) , ]
infoJL [ which ( infoJL$run == "3rd_2MP" ) , ]

table(infoJL$run)
     infoJL$run = gsub ( "2nd_A" , "2ndA" , infoJL$run )
     infoJL$run = gsub ( "2nd_B" , "2ndB" , infoJL$run )
     infoJL$run = gsub ( "3rd_1MP" , "3rd" , infoJL$run )
     infoJL$run = gsub ( "3rd_2MP" , "3rd" , infoJL$run )
     infoJL$rev = gsub ( "i0" , "" , infoJL$rev )


infoJL$composite = paste ( infoJL$run , infoJL$fwd , infoJL$rev
, sep = "_" )

##

table (table(infoJL$composite))
table (table (rownames(GG_JL_All_C)))

table (rownames(GG_JL_All_C) %in% infoJL$composite) ## 79 in
rownames(GG_JL_All_C) but not in  infoJL$composite
table (infoJL$composite %in% rownames(GG_JL_All_C) ) ## 22 in
infoJL$composite but not in rownames(GG_JL_All_C)
rownames(GG_JL_All_C) [ which(!rownames(GG_JL_All_C) %in%
infoJL$composite) ] ##

GG_JL_All_D = GG_JL_All_C [ which(rownames(GG_JL_All_C) %in%
infoJL$composite) , ]
GG_JL_All_BASE_D = GG_JL_All_BASE_C[
which(rownames(GG_JL_All_BASE_C) %in% infoJL$composite) , ]
infoJL2 = infoJL [ (infoJL$composite %in% rownames(GG_JL_All_C)
) , ]

## adding age and year
head(infoJL2)
     infoJL2$age = infoJL2$ID
     which(unlist (lapply ( strsplit ( infoJL2$age , split =
"_") , length))==3)
     infoJL2$age[1419:1421] = c(
"Uts_+3y_2015_1","Uts_+3y_2015_2","Uts_+3y_2015_3" )
     table(unlist (lapply ( strsplit ( infoJL2$age , split =
"_") , length)))

     infoJL2$year = unlist(lapply ( strsplit ( infoJL2$age ,
split = "_") , function(x) {x[3]} ))
     infoJL2$age = unlist(lapply ( strsplit ( infoJL2$age ,
split = "_") , function(x) {x[2]} ))
     infoJL2$age = gsub (" ","",infoJL2$age)
     infoJL2$age = gsub ( "Parent" ,"P",gsub ("y","",gsub
("\\+","",infoJL2$age)))
     table(infoJL2$age)
table ( rownames(GG_JL_All_D ) %in% infoJL2$composite)
table ( infoJL2$composite %in% rownames(GG_JL_All_D ) )
GG_JL_All_E = GG_JL_All_D [ match ( infoJL2$composite
,rownames(GG_JL_All_D ) ) , ]
GG_JL_All_BASE_E = GG_JL_All_BASE_D [ match ( infoJL2$composite
,rownames(GG_JL_All_BASE_D ) ) , ]
rownames(GG_JL_All_E ) == infoJL2$composite
```

```
rownames(GG_JL_All_BASE_E ) == infoJL2$composite
table(GG_JL_All_BASE_E[,7]) ## TT is 3SW six6
table(GG_JL_All_E[,7]) ## 0 us 3SW six6
table(GG_JL_All_BASE_E[,8]) ## CC is 3SW vgll3, AA=1SW
table(GG_JL_All_E[,8]) ## 0 us 3SW vgll3


### genoseuccess per individuals and loci

rownames(GG_JL_All_E )
(GG_JL_All_E )[1:10,1:10]
head(infoJL2)

sucIND = apply (  !is.na(GG_JL_All_E [,4:200]) , 1 , sum )  /
ncol (GG_JL_All_E)
sucLOC = apply (  !is.na(GG_JL_All_E [,4:200]) , 2 , sum )  /
nrow (GG_JL_All_E)

RUNN = unique (infoJL2$run)
run_range = t(sapply ( 1: length(RUNN) , function(i) { range (
which(infoJL2$run == RUNN[i]) ) }))

plot (sucIND )
abline(v=run_range[,1])
text (pos=4 , y=0.2, x = run_range[,1], labels = RUNN)
table(infoJL2$run , infoJL2$age)
head(infoJL2)
table (infoJL2 [which(infoJL2$run == "5th") ,"rev"] )
table (infoJL2 [which(infoJL2$run == "6th") ,"rev"] )
table (infoJL2 [which(infoJL2$run == "2ndA") ,"rev"] )
table (infoJL2 [which(infoJL2$run == "2ndB") ,"rev"] )

##
write.table ( GG_JL_All_E , file = "GG_JL_All_E.txt"  , quote =
F, sep= "\t" )
write.table ( infoJL2 , file = "infoJL2.txt"  , quote = F, sep=
"\t" )

### Jan laine additional genotypings ##

# user IT1-16-System Gen Seq Temp Rerun BS17-18-19 and JanMix
# this is the rerun of IT1-114 (BS) and IT1-109 (Jan).

# IT1-109 run in jan alines had particularly low genotyping
success, which we re-tun it in "IT1-16".

## running and genotyoing baltic genos
## tutku aykanat, 16.09.17

setwd("C:/Users/tutayk/Desktop/JanLaine")


## MERGING duplicate IDs in "IT1-109" and "IT1-16".
{
     ## load both "IT1-109" and "IT1-16"

     load ( file = "C:/Users/tutayk/Desktop/JanLaine/user IT1-
16/IT1-16_COVSpreTRIAL" )
     load ( file =
"C:/Users/tutayk/Desktop/JanLaine/5th_COVSpreTRIAL")

     PRECOVFILE1 = get( "5th_COVSpreTRIAL")
     PRECOVFILE2 = get( "IT1-16_COVSpreTRIAL") #Includes also
```

```
baltic genos. Exclude them before combining

     dim (PRECOVFILE1)
     dim (PRECOVFILE2)

     colnames (PRECOVFILE1)
     colnames (PRECOVFILE2)

     which ( !colnames (PRECOVFILE1) %in% colnames (PRECOVFILE2)
) ## only one is missing in teh initial run as oppose to re-run.
     PRECOVFILE1[,32] ## essentially nothing here.

     PRECOVFILE1b = PRECOVFILE1[,-32]
     PRECOVFILE2b = PRECOVFILE2 [ ,  match (
colnames(PRECOVFILE1b) , colnames(PRECOVFILE2) )]

     dim(PRECOVFILE1b)
     dim(PRECOVFILE2b)

     table ( colnames(PRECOVFILE1b) == colnames(PRECOVFILE2b) )

     ##merging datasets here
     PRECOVFILE = t(sapply ( 1:nrow(PRECOVFILE1b) , function(j)
{ print(j)
               sapply ( 1:ncol(PRECOVFILE1b) , function(i) {
               A1 = c(
PRECOVFILE1b[j,i][[1]],PRECOVFILE2b[j,i][[1]] )
               ANa = as.numeric(c(
PRECOVFILE1b[j,i][[1]],PRECOVFILE2b[j,i][[1]] ))
               AN = sum (ANa , na.rm=T)
               AF = list (c( A1[is.na(ANa)] , AN )     )
               names(AF) = names(PRECOVFILE1b[j,i])
               AF
          })
               }))

     ###gebnotyoing merged datasets
               PRECOVFILE = PRECOVFILE
               lociBYlociPlot = T # a boolean by which you
decide to have these plots or not.
               THR = 10 # the thrshold which a genotype will not
be called for an individuals if coverage is less than this.
               THRmvn = 15 # same as above, but defined for MVN
plot. Note that we keep mvn tehrhold a bit higher.
               folder.to.create = paste(
"C:/Users/tutayk/Desktop/JanLaine/6threrun_combined",
"/genotyping.resultsANDfigures", sep="") # a folder will be
cretaed by "genotyping" function, in which results will be
stored.
               genotyping ( PRECOVFILE, lociBYlociPlot, THR ,
THRmvn, folder.to.create)


     ### adding new run and re-runs to the frame (F for final)

     IT16_IT109_num = t ( read.table ( stringsAsFactors = F, sep
= "\t" , h=T ,
"C:/Users/tutayk/Desktop/JanLaine/6threrun_combined/genotyping.r
esultsANDfigures/genotype.by.numbers.txt" ))
     IT16_IT109_base  = t ( read.table ( stringsAsFactors = F,
sep = "\t" , h=T ,
"C:/Users/tutayk/Desktop/JanLaine/6threrun_combined/genotyping.r
esultsANDfigures/genotype.by.bases.txt"))
```

```
     dim(IT16_IT109_num)
     dim(IT16_IT109_base)
     rownames(IT16_IT109_base) = gsub ( "X" , "5th_" ,
rownames(IT16_IT109_base) )
     compID_rerun = rownames(IT16_IT109_base)
     table(compID_rerun %in% infoJL2$composite)

     ## frames
     dim(GG_JL_All_E)
     dim(GG_JL_All_BASE_E)
     dim(infoJL2)
     head(infoJL2)

     ## run is same as 5th

     IT16_IT109_num [1:10,1:10]

     rerun_ind = which(infoJL2$run == "5th")
     infoJL_rerun = infoJL2 [ rerun_ind , ]
     infoJL_rerun [ which ( !infoJL_rerun$composite %in%
compID_rerun ), "ID"]


     ### exclude "Uts_+3y_2015_37" from master file. Note that
it is not genotyped in there as well
     GG_JL_All_Eb = GG_JL_All_E [ -which ( infoJL2$ID ==
"Uts_+3y_2015_37" ) , ]
     GG_JL_All_BASE_Eb = GG_JL_All_BASE_E [ -which ( infoJL2$ID
== "Uts_+3y_2015_37" ) , ]
     infoJL3 = infoJL2 [ -which ( infoJL2$ID ==
"Uts_+3y_2015_37" ) , ]

     ## some confrimation on the concordance between genotypes

          dim(IT16_IT109_num)
          dim(IT16_IT109_base)
          table(compID_rerun %in% infoJL3$composite)

          match_index1 = match ( compID_rerun ,
infoJL3$composite )

          table ( IT16_IT109_num[,-6] ==  ( GG_JL_All_Eb [
match_index1 , -c(1:3,9) ]  ) , useNA = "a" )
          table ( IT16_IT109_base[,-6] ==  ( GG_JL_All_BASE_Eb [
match_index1 , -c(1:3,9) ]  ) , useNA = "a" )

          table(IT16_IT109_num[,-6],useNA="a")
          table(GG_JL_All_Eb [ match_index1 , -c(1:3,9) ]
,useNA="a")
     }
## ENDS: MERGING duplicate IDs in "IT1-109" and "IT1-16".

## NOW MERGING
GG_JL_All_Eb [ match_index1 , -c(1:3) ]  =  IT16_IT109_num
GG_JL_All_BASE_Eb [ match_index1 , -c(1:3) ]  =  IT16_IT109_base

     #marke -9 with NA
     GG_JL_All_Eb [GG_JL_All_Eb == -9  ] = NA
     GG_JL_All_BASE_Eb [GG_JL_All_BASE_Eb == -9  ] = NA

## concordance with 1MP vs 3rd dup data
{
```

```
     dupID = names ( which ( table(infoJL3[,1]) == 2 ) )
     dupID

     DD = t ( sapply( 1 : length(dupID) , function(i) { print(i)
          KK = GG_JL_All_Eb [ which (  infoJL3[,1] %in% dupID[i]
) , -c(1:3) ]
          c ( length(which(KK[1,] == KK [2,])) ,
length(which(KK[1,] != KK [2,])) )
          }) )

     sum(DD[,1]) / sum(DD) ## %99.4 concordance.
}
## ENDS: concordance with 1MP vs 3rd dup data

##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
~~~~~~~~~~~~~~~~~~~~~~##
## MERGE info taht has two IDs ## (these are as teh result of
1mp vs 2MP comparison) ##
## 1 MP vs 2MP comparsion were made using the same IDs in 3rd
run ~~~~~~~~~~~~~~~~~~~##
## duplicate IDs have the same fwd barcode (reverse 1-4 and 2-3
goes together) ~~~~~~##
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
~~~~~~~~~~~~~~~~~~~~~~##

# note that rev barcode 1-4 goes btw 1:86 (fwd barcode) and 2-3
btw (1-70).
# other IDS in this run is irrelevant

load ( file =
"C:/Users/tutayk/Desktop/JanLaine/3rd_COVSpreTRIAL")
{
     table ( table(infoJL3$ID) )
     dupIDs = names(which ( table(infoJL3$ID)==2 ))

     # 3rd ID 1&4 and 2&3 rev barcodes are duplicates
     t(sapply ( 1:length(dupIDs) , function(i) {
     ID_ind = which ( infoJL3$ID %in% dupIDs[i] )
          c ( infoJL3 [ ID_ind[1] , c ("run" ,"fwd" , "rev") ] ,
infoJL3 [ ID_ind[2] , c ("run" ,"fwd" , "rev") ]  )
          }))
     infoJL3 [ infoJL3$run == "3rd" ,2:4 ]

     PRECOVFILE1 = get( "3rd_COVSpreTRIAL")

     #combined rev barcodes 1-4 and 2-3

     PRECOVFILE1b = PRECOVFILE1 [ , grep ( "_1" ,
colnames(PRECOVFILE1) )  ]
     PRECOVFILE2b = PRECOVFILE1 [ , grep ( "_2" ,
colnames(PRECOVFILE1) )  ]
     PRECOVFILE3b_pre = PRECOVFILE1 [ , grep ( "_3" ,
colnames(PRECOVFILE1) )  ]
     PRECOVFILE4b = PRECOVFILE1 [ , grep ( "_4" ,
colnames(PRECOVFILE1) )  ]

     PRECOVFILE3b = cbind ( PRECOVFILE3b_pre[,1:28],0,
PRECOVFILE3b_pre[,29:95])
     colnames(PRECOVFILE3b)[29] = "29_3"

     PRECOVFILE14 = t(sapply ( 1:nrow(PRECOVFILE1b) ,
function(j) { print(j)
               sapply ( 1:ncol(PRECOVFILE1b) , function(i) {
```

```
                A1 = c(
PRECOVFILE1b[j,i][[1]],PRECOVFILE4b[j,i][[1]] )
                ANa = as.numeric(c(
PRECOVFILE1b[j,i][[1]],PRECOVFILE4b[j,i][[1]] ))
                AN = sum (ANa , na.rm=T)
                AF = list (c( A1[is.na(ANa)] , AN )     )
                names(AF) = names(PRECOVFILE1b[j,i])
                AF
          })
                }))[,1:86]
     PRECOVFILE23 = t(sapply ( 1:nrow(PRECOVFILE2b) ,
function(j) { print(j)
                sapply ( 1:ncol(PRECOVFILE2b) , function(i) {
                A1 = c(
PRECOVFILE2b[j,i][[1]],PRECOVFILE3b[j,i][[1]] )
                ANa = as.numeric(c(
PRECOVFILE2b[j,i][[1]],PRECOVFILE3b[j,i][[1]] ))
                AN = sum (ANa , na.rm=T)
                AF = list (c( A1[is.na(ANa)] , AN )     )
                names(AF) = names(PRECOVFILE2b[j,i])
                AF
          })
                }))[,1:70]

     ##
     PRECOVFILE_F = cbind ( PRECOVFILE14 , PRECOVFILE23 )

     ###genotyping merged datasets

          PRECOVFILE = PRECOVFILE_F
          lociBYlociPlot = T # a boolean by which you decide to
have these plots or not.
          THR = 10 # the thrshold which a genotype will not be
called for an individuals if coverage is less than this.
          THRmvn = 15 # same as above, but defined for MVN plot.
Note that we keep mvn tehrhold a bit higher.
          folder.to.create = paste(
"C:/Users/tutayk/Desktop/JanLaine/3rd_dups combined",
"/genotyping.resultsANDfigures", sep="") # a folder will be
cretaed by "genotyping" function, in which results will be
stored.
          genotyping ( PRECOVFILE, lociBYlociPlot, THR , THRmvn,
folder.to.create)


     ### adding new run and re-runs to the frame (F for final)

     run_3rd_dups = t ( read.table ( stringsAsFactors = F, sep =
"\t" , h=T , "C:/Users/tutayk/Desktop/JanLaine/3rd_dups
combined/genotyping.resultsANDfigures/genotype.by.numbers.txt"
))
     run_3rd_dups_base  = t ( read.table ( stringsAsFactors = F,
sep = "\t" , h=T , "C:/Users/tutayk/Desktop/JanLaine/3rd_dups
combined/genotyping.resultsANDfigures/genotype.by.bases.txt"))
     rownames(run_3rd_dups_base) = gsub ( "X" , "3rd_" ,
rownames(run_3rd_dups_base) )
     compID_rerun = rownames(run_3rd_dups_base)
     table(compID_rerun %in% infoJL2$composite)

     ## frames
     dim ( GG_JL_All_Eb )
     dim ( GG_JL_All_BASE_Eb )
     dim ( infoJL3 )
```

```
     ## first, remove 3rd run rev barcode 4 and 3 (we will keep
their dup info 1 and 2, repscetively)
     remIND3 = grep ( "3rd_.{1,2}_3",rownames(GG_JL_All_Eb) )
     remIND4 = grep ( "3rd_.{1,2}_4",rownames(GG_JL_All_Eb) )
     remIND_3rdrun = c(remIND3,remIND4)

     GG_JL_All_Ec = GG_JL_All_Eb [-remIND_3rdrun,]
     GG_JL_All_BASE_Ec = GG_JL_All_BASE_Eb [-remIND_3rdrun,]
     infoJL3b = infoJL3 [-remIND_3rdrun,]

     # now adding new genotype framae to 3rd run 1 and 2nd rev
barcodes

     merge_ind = which(infoJL3b$run == "3rd")
     infoJL_rerun = infoJL3b [ merge_ind , ]

     ## some confrimation on the concordance between genotypes

          dim(run_3rd_dups)
          dim(run_3rd_dups_base)
          table(compID_rerun %in% infoJL3b$composite)

          match_index1 = match ( compID_rerun ,
infoJL3b$composite )

          # only 15 genotyped changed.
          table ( run_3rd_dups[,-6] ==  ( GG_JL_All_Ec [
match_index1 , -c(1:3,9) ]  ) , useNA = "a" )
          table ( run_3rd_dups_base[,-6] ==  ( GG_JL_All_BASE_Ec
[ match_index1 , -c(1:3,9) ]  ) , useNA = "a" )

          GG_JL_All_Ec [ match_index1 , -c(1:3) ] = run_3rd_dups
          GG_JL_All_BASE_Ec [ match_index1 , -c(1:3) ] =
run_3rd_dups_base

          table(GG_JL_All_Ec[,-c(1:3,9)],useNA="a")
          table(GG_JL_All_BASE_Ec[,-c(1:3,9)],useNA="a")

          #marke -9 with NA
          GG_JL_All_Ec [GG_JL_All_Ec == -9  ] = NA
          GG_JL_All_BASE_Ec [GG_JL_All_BASE_Ec == -9  ] = NA
}

##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
~~~~~~~~~~~~~~~~~~~~~~~~~~##
## ENDS: MERGE info taht has two IDs ## (these are as teh result
of 1mp vs 2MP comparison) ##
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
~~~~~~~~~~~~~~~~~~~~~~~~~~##

##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~##
#~~~~~~ FINAL DATA FARMES FOR ANLAYSIS ~~~~~~#
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~##

     dim(GG_JL_All_Ec)
     dim(GG_JL_All_BASE_Ec)
     dim(infoJL3b)
     head(infoJL3b)

##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~##
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~##
```

```
setwd("C:/Users/tutayk/Desktop/JanLaine")

### genoseuccess per individuals and loci

rownames(GG_JL_All_Ec )
(GG_JL_All_E )[1:10,1:10]
head(infoJL2)

sucIND = apply (  !is.na(GG_JL_All_Ec [,4:200]) , 1 , sum )  /
197
sucLOC = apply (  !is.na(GG_JL_All_Ec [,4:200]) , 2 , sum )  /
nrow (GG_JL_All_Ec)

RUNN = unique (infoJL3b$run)
run_range = t(sapply ( 1: length(RUNN) , function(i) { range (
which(infoJL3b$run == RUNN[i]) ) }))



plot (sucIND )
abline(v=run_range[,1])
text (pos=4 , y=0.2, x = run_range[,1], labels = RUNN)

table(infoJL3b$run , infoJL3b$age)


## vgll3 and sic6 success per RUN
suc2 = t(sapply(1: length(RUNN) , function(i) {
     KK = which (infoJL3b$run == RUNN[i])
     KK2 = !is.na(GG_JL_All_Ec [KK,4:200])
     all = sum(KK2)/length(c(KK2))
     vgll3 = sum(!is.na(GG_JL_All_Ec [KK,8])) / length(KK)
     six6 = sum(!is.na(GG_JL_All_Ec [KK,7])) / length(KK)
     c(all,vgll3,six6)

}))

rownames(suc2) = RUNN
colnames(suc2) = c("all","vgll3","six6")

write.table ( file = "genotype.success_afterReRON5th.txt" , suc2
, quote = F, sep= "\t" )
getwd()

##
getwd()
write.table ( GG_JL_All_Ec , file = "GG_JL_All_Ec.txt"  , quote
= F, sep= "\t" )
write.table ( infoJL3b , file = "infoJL3b.txt"  , quote = F,
sep= "\t" )

table ( GG_JL_All_Ec[,4] , GG_JL_All_Ec[,8] )
table ( GG_JL_All_BASE_Ec[,4] , GG_JL_All_BASE_Ec[,8] )

##
##
##
##


## geno suc per loci (excluding loci that has less than 80%
success except vgll3 amd akap11)
sucLOC = apply ( !is.na(GG_JL_All_Ec[,4:200]) , 2 , sum) /
```

```
nrow(GG_JL_All_Ec)
plot(sucLOC)
sucLOC[1:10]
abline(h=0.80)
EXloc = which(sucLOC<0.789) ## using 0.789 insetad of 80 only
excludes vgll3 additionally.
length(EXloc)
runss


# excluded loci by eye (based on topology)
Exloc2 = c ( 20 , 78 , 89 , 93 , 116 , 199 , 141 , 155 , 157 ,
182 , 191 )

## zero success per loci in 5th run (except akap11)
EXloc5th = which(apply ( !is.na(GG_JL_All_Ec [
which(infoJL3b$run == "5th") ,-c(1:3) ] ) , 2 , mean )< 0.01)[-
1]
EXlocALL = unique(c(EXloc5th , EXloc , Exloc2) )+3
##
     GG_JL_All_Fpre = GG_JL_All_Ec [  , -EXlocALL]
     GG_JL_All_BASE_Fpre = GG_JL_All_BASE_Ec[  , -EXlocALL]

##
## geno suc per ind

sucIND = apply ( !is.na(GG_JL_All_Fpre[ ,-c(1:3)]) , 1 , sum) /
ncol (GG_JL_All_Fpre[ ,-c(1:3)])
plot(sucIND)
abline(h=0.7)

#which ( sucIND < 0.8 & infoJL3b$age == "P" ) # keep one parent
with slightly low genotyping success
EXind = which ( sucIND < 0.7 & infoJL3b$age != "P" )

GG_JL_All_F = GG_JL_All_Fpre [ -EXind , ]
GG_JL_All_BASE_F = GG_JL_All_BASE_Fpre[ -EXind , ]
infoJL4 = infoJL3b [ -EXind , ]

dim(GG_JL_All_F)
dim(GG_JL_All_BASE_F)
dim(infoJL4)

### AF diffrece
GG_JL_All_G =  apply ( GG_JL_All_F [,-c(1:3,9)], 2, as.numeric )
GG_JL_All_BASE_G = GG_JL_All_BASE_F [,-c(1:3,9)]

     #some run soceific threholds
by ( as.numeric(GG_JL_All_F[,9]) , infoJL4$run , mean , na.rm=T
) # 5th run has high sdy ratios
runss =  unique (infoJL4$run)

i=9
ind1 = which ( infoJL4$run == runss[i] )
plot(as.numeric(GG_JL_All_F[ind1,9]));abline(h=0.1,col="red");ab
line(h=0.5,col="red")
thr_runSpec = list
(c(0.1,0.2),c(0.1,0.2),c(0.1,0.2),c(0.3,0.5),c(0.3,0.5),c(0.5,2)
,c(0.3,0.5),c(0.1,0.5),c(0.1,0.2))
for (i in 1:9) { print(i)
     ind_male = which ( infoJL4$run == runss[i] &
as.numeric(GG_JL_All_F[,9]) > thr_runSpec[[i]][2] )
     ind_female = which ( infoJL4$run == runss[i] &
```

```
as.numeric(GG_JL_All_F[,9]) < thr_runSpec[[i]][1] )

     infoJL4$sexG[ind_male] = "M"
     infoJL4$sexG[ind_female] = "F"
}

plot ( as.numeric(GG_JL_All_F[,9]) , col =
as.numeric(as.factor(infoJL4$sexG))+1 )
plot ( as.numeric(GG_JL_All_F[,9]) , col =
as.numeric(as.factor(infoJL4$sexG))+1 , ylim=c(0,10))
infoJL4$year = as.numeric(infoJL4$year)
infoJL4$age = as.numeric(infoJL4$age)

## SNPs remain in the anlaysis as a function of category
colnames(SET_FINAL)
AA = SET_FINAL [ match ( colnames(GG_JL_All_G) , SET_FINAL$id )
, "cat" ]
AA2= AA
AA2 [grep("ull",AA) ] = "null"
AA2 [grep("ut",AA) ] = "out"
AA3 = as.numeric(as.factor(AA2))
#
## adding sea age infromation
sea_ageinfo = read.table (stringsAsFactors = F ,
"C:/Users/tutayk/Desktop/JanLaine/Parent_2011_Sample_information
.txt" , h=T )
p_index = which(is.na(infoJL4$age))

as.numeric(gsub ( "Uts_Parent_2011_", "" , infoJL4$ID[p_index] )
) == sea_ageinfo[,1]

infoJL4$age[p_index] = paste ( "SW" , sea_ageinfo$SA , sep="" )
infoJL4$sexP = NA
infoJL4$W = NA
infoJL4$L = NA
infoJL4$sexP[p_index] = sea_ageinfo$Sex
infoJL4$W[p_index] = sea_ageinfo$W
infoJL4$L[p_index] = sea_ageinfo$L
table(infoJL4$sexP == infoJL4$sexG) ## all true

## adding offspring info
     infoJL4$offspring =  NA
          JLped = read.table (stringsAsFactors = F ,
"C:/Users/tutayk/Desktop/JanLaine/Pedigree_Adult2011_offspring20
12.txt" , h=T )
          JLped$Dam[ JLped$Dam == "us"]=NA
          JLped$Sire[ JLped$Sire == "us"]=NA

     JLped$OffspID2 = paste ( substr(JLped$OffspID,1,13) ,
as.numeric(substr(JLped$OffspID,14,16)) , sep = "" )

##
par_index = grep("SW" , infoJL4$age )
par.w.off = table (na.exclude( c(JLped$Dam , JLped$Sire ) ) )
names(par.w.off) =
as.numeric(gsub("Uts_11A_","",names(par.w.off)))

for( i in 1:54) {
     A  = which ( names(par.w.off)  %in% i )
     if ( length(A) == 0 )      infoJL4 [ par_index[i] ,
"offspring"] = 0 else {infoJL4 [ par_index[i] , "offspring"] =
par.w.off[A] }
     }
```

```
infoJL4 [ par_index, "offspring"]
##

infoJL4$parDam = NA
infoJL4$parSire = NA

table ( JLped$OffspID2  %in% infoJL4$ID )
JLped2 = JLped [ ( JLped$OffspID2  %in% infoJL4$ID ) ,  ]

infoJL4$parDam [ match ( JLped2$OffspID2 , infoJL4$ID ) ] =
JLped2$Dam
infoJL4$parSire [ match ( JLped2$OffspID2 , infoJL4$ID ) ] =
JLped2$Sire

infoJL4$age2 = infoJL4$age
infoJL4$age2 [grep("SW" , infoJL4$age )] = "P"

### ADDING YAnn\s adults to teh dataset ##
{
     #note adults numbers doesnt match (make a cross in teh plot
:)
     yannAdults_num = read.table ( h=T, stringsAsFactors = F ,
sep = "\t" ,
"C:/Users/tutayk/Desktop/JanLaine/mainstem_2011_adults.txt" )
     yannAdults_base = read.table ( h=T, stringsAsFactors = F ,
sep = "\t" ,
"C:/Users/tutayk/Desktop/JanLaine/Spawners_2011_tutku.txt" )

     colnames(yannAdults_base)  = gsub ( "X" , "" , gsub (
"\\.." ,  "" , colnames(yannAdults_base) ) )
     colnames(GG_JL_All_G) %in% colnames(yannAdults_base)
     yannSNPind = match ( colnames(GG_JL_All_G) ,
colnames(yannAdults_base)  )
     GG_JL_All_BASE_G[1:5,1:15]

     ## geno succes yann (filter <0.7 succes ind.)
     excldYANN = which ( 1- apply ( is.na(yannAdults_base
[,yannSNPind]) , 1 , mean )  < 0.7 )

     GG_JL_All_BASE_Gy = (  rbind ( GG_JL_All_BASE_G ,
yannAdults_base [-excldYANN,yannSNPind] )  )
     GG_JL_All_BASE_Gy = apply(GG_JL_All_BASE_Gy,2,as.character)
     yannAdults_base2 = yannAdults_base [-excldYANN,yannSNPind]
     dim(yannAdults_base2)

     yannAdults_base2
     dim(yannAdults_base2)
     yannAdults_num2 = yannAdults_base2
     ## matcheding genotypes based on letters (bases) amd then
generating a data file with numbers
     match_ind_Y = t( sapply (1:ncol(GG_JL_All_BASE_G),
function(i) {
               KK = c(0,1,2)
               BB = GG_JL_All_BASE_G[,i] [ match ( KK ,
GG_JL_All_G[,i] ) ]
               names(BB) = KK
               BB
               }) )
          for ( i in 1:ncol(yannAdults_base2) )  {
               yannAdults_num2[,i] = as.numeric (
names(match_ind_Y[i,]) [ match ( yannAdults_base2[,i] ,
match_ind_Y[i,] ) ] )
          }
```

```
          dim(match_ind)
          yannAdults_num2[1:10,1:10]
          yannAdults_num2 =  apply ( yannAdults_num2 , 2,
as.numeric)
          #check if numbers match
          plot ( apply ( yannAdults_num2  , 2, mean, na.rm= T )
, apply ( GG_JL_All_G  , 2, mean, na.rm= T )  )
          abline(0,1)
          points ( pch=20,col="red",apply ( yannAdults_num2  ,
2, mean, na.rm= T ) [c( 1:5 , 166 : 167 ) ], apply ( GG_JL_All_G
, 2, mean, na.rm= T ) [c( 1:5 , 166 : 167 ) ] )
          GG_JL_All_Gy = rbind ( GG_JL_All_G , yannAdults_num2 )
          GG_JL_All_BASE_Gy
          yann_phen = as.data.frame(matrix ( NA , nrow =
nrow(yannAdults_num2) , ncol (infoJL4) ))
          names(yann_phen) = names (infoJL4)
          yannAdults_base [ ,
colnames(yannAdults_base)[c(1:2,199:214)]  ]
     #GENOSEX OF YANNS (same as phenosex)
          yannAdults_base [ ,
colnames(yannAdults_base)[c(1:2,199:214)]  ]
          plot ( ylim=c(0,4),yannAdults_base [ , "aveSDY" ]  ,
pch=20 , col = as.numeric(as.factor(yannAdults_base [ , "sex"
])))
          abline(h=c(0.1,0.3))
          sexG_yann = gsub ( "male", "M" , gsub ( "female", "F"
, yannAdults_base_phen  [ , "sex" ] ))
          yannAdults_base_phen = yannAdults_base [-excldYANN,]
          yannAdults_base_phen  [ 1:3,
colnames(yannAdults_base)[c(1:2,199:214)]  ]
               yann_phen$ID = yannAdults_base_phen$ID
               yann_phen$sexP = sexG_yann # bote above sexP and
sexG is smae
               yann_phen$sexG = sexG_yann
               yann_phen$year = 2011
               yann_phen$age2 = "Py"
               yann_phen$composite = yannAdults_base_phen $num
               yann_phen$age = yannAdults_base_phen $history2
                    yann_phen$age [nchar(yann_phen$age)==1] =
paste ( "SW" , yann_phen$age [nchar(yann_phen$age)==1]  , sep =
"" )

          #adding length and weight info
                              yannAdults_LW = read.table ( h=T,
stringsAsFactors = F , sep = "\t" ,
"C:/Users/tutayk/Desktop/PS/catches_last.txt" )
          dim(yannAdults_LW)
          head(yannAdults_LW)
          head(yann_phen)
          yann_phen$L = as.numeric( gsub ( "," , "." ,
yannAdults_LW [  match ( yann_phen$composite , yannAdults_LW$num
) , "length" ] ) )
          yann_phen$W = as.numeric( gsub ( "," , "." ,
yannAdults_LW [  match ( yann_phen$composite , yannAdults_LW$num
) , "weight" ] ) )
          infoJL4y = rbind ( infoJL4 , yann_phen )
}
### ENDS: ADDING YAnn\s adults to teh dataset ##
# hierfstat (R package) format
{
     dim(GG_JL_All_Gy )
     JL_hier_pre1 =matrix(NA, nrow=2*nrow(GG_JL_All_Gy ) , ncol
= ncol(GG_JL_All_Gy ))
```

```
          for ( i in 1:nrow(GG_JL_All_Gy) ) {
               AA = GG_JL_All_Gy[i,]
                    AA[AA==1] = "1,2"
                    AA[AA==0] = "1,1"
                    AA[AA==2] = "2,2"
                    AA[is.na(AA)] = "9,9"
               JL_hier_pre1 [ (i*2-1):(i*2) , ] = rbind (
substr(AA,1,1) ,substr(AA,3,3) )
          }
          JL_hier_pre1 [JL_hier_pre1== 9]=NA
     ID = rep ( 1:nrow(GG_JL_All_Gy )   , each =2 )
     AGE = rep ( infoJL4y$age   , each =2 )
     SEX = rep ( infoJL4y$sexG   , each =2 )

     JL_hier_pre2 = cbind ( ID , AGE , SEX , JL_hier_pre1 )
     JL_hier_pre3 = as.data.frame (JL_hier_pre2)
     for ( i in 4:ncol(JL_hier_pre3)  ) {JL_hier_pre3[,i] =
as.numeric(JL_hier_pre3[,i])}
     JL_hier_pre3[,1] = as.numeric(JL_hier_pre3[,1])
     str(JL_hier_pre3 )
     JL_hier_pre3$AGE = as.character (JL_hier_pre3$AGE)
     JL_hier_pre3 [ grep ( "SW" , JL_hier_pre3$AGE) , "AGE" ] =
4
     JL_hier_pre3$AGE = as.numeric(JL_hier_pre3$AGE)+1
     table(JL_hier_pre3$AGE)
     JL_hier = JL_hier_pre3
}
# ENDS: hierfstat format
## genetics fromat
{
library(genetics)
     JL_genotype = sapply ( 1: ncol (GG_JL_All_BASE_Gy) ,
simplify = F , function(i) { genotype (GG_JL_All_BASE_Gy[,i] ,
sep = "" ) })
     JL_genotype_age = sapply ( 1:  ncol (GG_JL_All_BASE_Gy) ,
simplify = F, function(j) { print(j)
                              sapply (
sort(unique(infoJL4y$age2)) , simplify = F, function(i) {
                                   if (
sum(!is.na(GG_JL_All_BASE_Gy[infoJL4y$age2 == i,j])) == 0 ) NA
else {
                                        genotype
(GG_JL_All_BASE_Gy[infoJL4y$age2 == i,j] , sep = "" ) }
                                   })
                         })
}
## ENDS: genetics (R package) format
save ( infoJL4y, file ="infoJL4y" )
save ( GG_JL_All_Gy, file ="GG_JL_All_Gy" )
save ( GG_JL_All_BASE_Gy, file ="GG_JL_All_BASE_Gy" )
save ( JL_hier, file = "JL_hier" )
save ( JL_genotype, file = "JL_genotype" )
save ( JL_genotype_age, file = "JL_genotype_age" )
```

APPENDIX J: The R script for Fisher's exact test for genotype frequencies of all SNP containing loci between different age group and sexes

```
#By age and sex

#Males
zeroM <- GG_JL_All_Gy[which(infoJL4y$age3 %in% 0 & infoJL4y$sexG
== "M"),]
firstM <- GG_JL_All_Gy[which(infoJL4y$age3 %in% 1 &
infoJL4y$sexG == "M"),]
secondM <- GG_JL_All_Gy[which(infoJL4y$age3 %in% 2 &
infoJL4y$sexG == "M"),]
thirdM <- GG_JL_All_Gy[which(infoJL4y$age3 %in% 3 &
infoJL4y$sexG == "M"),]
parentM <- GG_JL_All_Gy[which(infoJL4y$age3 %in% "P" &
infoJL4y$sexG == "M"),]

#Females
zeroF <- GG_JL_All_Gy[which(infoJL4y$age3 %in% 0 & infoJL4y$sexG
== "F"),]
firstF <- GG_JL_All_Gy[which(infoJL4y$age3 %in% 1 &
infoJL4y$sexG == "F"),]
secondF <- GG_JL_All_Gy[which(infoJL4y$age3 %in% 2 &
infoJL4y$sexG == "F"),]
thirdF <- GG_JL_All_Gy[which(infoJL4y$age3 %in% 3 &
infoJL4y$sexG == "F"),]
parentF <- GG_JL_All_Gy[which(infoJL4y$age3 %in% "P" &
infoJL4y$sexG == "F"),]

### -Age group fisher exact- ###
#0y vs parents (No -log)
zeroVSparent <- sapply ( 1: ncol(GG_JL_All_Gy) , function(i) {
print(i)
  group.genotypes.A <- zero[,i]
  group.genotypes.B <- parent[,i]
  genotype.freqs.A<-c(sum(group.genotypes.A %in% 0),
sum(group.genotypes.A %in% 1), sum(group.genotypes.A %in% 2))
  genotype.freqs.B<-c(sum(group.genotypes.B %in% 0),
sum(group.genotypes.B %in% 1), sum(group.genotypes.B %in% 2))
  geno.comp. <- cbind(genotype.freqs.A,genotype.freqs.B)
  if (sum(geno.comp.)<1) NA else {
    #fisher.test (  cbind ( compA1 , compB1 ) )
    (fisher.test (  geno.comp. )$p) }
} )
names(zeroVSparent)<-colnames(GG_JL_All_Gy)

#0y vs parentM (No -log)
zeroVSparentM <- sapply ( 1: ncol(GG_JL_All_Gy) , function(i) {
print(i)
  group.genotypes.A <- zero[,i]
  group.genotypes.B <- parentM[,i]
  genotype.freqs.A<-c(sum(group.genotypes.A %in% 0),
sum(group.genotypes.A %in% 1), sum(group.genotypes.A %in% 2))
  genotype.freqs.B<-c(sum(group.genotypes.B %in% 0),
sum(group.genotypes.B %in% 1), sum(group.genotypes.B %in% 2))
  geno.comp. <- cbind(genotype.freqs.A,genotype.freqs.B)
  if (sum(geno.comp.)<1) NA else {
    #fisher.test (  cbind ( compA1 , compB1 ) )
    (fisher.test (  geno.comp. )$p) }
} )
names(zeroVSparentM)<-colnames(GG_JL_All_Gy)
```

```
#0y vs parentF (No -log)
zeroVSparentF <- sapply ( 1: ncol(GG_JL_All_Gy) , function(i) {
print(i)
  group.genotypes.A <- zero[,i]
  group.genotypes.B <- parentF[,i]
  genotype.freqs.A<-c(sum(group.genotypes.A %in% 0),
sum(group.genotypes.A %in% 1), sum(group.genotypes.A %in% 2))
  genotype.freqs.B<-c(sum(group.genotypes.B %in% 0),
sum(group.genotypes.B %in% 1), sum(group.genotypes.B %in% 2))
  geno.comp. <- cbind(genotype.freqs.A,genotype.freqs.B)
  if (sum(geno.comp.)<1) NA else {
    #fisher.test (  cbind ( compA1 , compB1 ) )
    (fisher.test (  geno.comp. )$p) }
} )
names(zeroVSparentF)<-colnames(GG_JL_All_Gy)

#parentM VS parentF (No -log)
parentMVSparentF <- sapply ( 1: ncol(GG_JL_All_Gy) , function(i)
{ print(i)
  group.genotypes.A <- parentM[,i]
  group.genotypes.B <- parentF[,i]
  genotype.freqs.A<-c(sum(group.genotypes.A %in% 0),
sum(group.genotypes.A %in% 1), sum(group.genotypes.A %in% 2))
  genotype.freqs.B<-c(sum(group.genotypes.B %in% 0),
sum(group.genotypes.B %in% 1), sum(group.genotypes.B %in% 2))
  geno.comp. <- cbind(genotype.freqs.A,genotype.freqs.B)
  if (sum(geno.comp.)<1) NA else {
    #fisher.test (  cbind ( compA1 , compB1 ) )
    (fisher.test (  geno.comp. )$p) }
} )
names(parentMVSparentF)<-colnames(GG_JL_All_Gy)


#0y vs 1y (No -log)
zeroVSfirst <- sapply ( 1: ncol(GG_JL_All_Gy) , function(i) {
print(i)
  group.genotypes.A <- zero[,i]
  group.genotypes.B <- first[,i]
  genotype.freqs.A<-c(sum(group.genotypes.A %in% 0),
sum(group.genotypes.A %in% 1), sum(group.genotypes.A %in% 2))
  genotype.freqs.B<-c(sum(group.genotypes.B %in% 0),
sum(group.genotypes.B %in% 1), sum(group.genotypes.B %in% 2))
  geno.comp. <- cbind(genotype.freqs.A,genotype.freqs.B)
  if (sum(geno.comp.)<1) NA else {
    #fisher.test (  cbind ( compA1 , compB1 ) )
    (fisher.test (  geno.comp. )$p) }
} )
names(zeroVSfirst)<-colnames(GG_JL_All_Gy)


#0y vs 2y (No -log)
zeroVSsecond <- sapply ( 1: ncol(GG_JL_All_Gy) , function(i) {
print(i)
  group.genotypes.A <- zero[,i]
  group.genotypes.B <- second[,i]
  genotype.freqs.A<-c(sum(group.genotypes.A %in% 0),
sum(group.genotypes.A %in% 1), sum(group.genotypes.A %in% 2))
  genotype.freqs.B<-c(sum(group.genotypes.B %in% 0),
sum(group.genotypes.B %in% 1), sum(group.genotypes.B %in% 2))
  geno.comp. <- cbind(genotype.freqs.A,genotype.freqs.B)
  if (sum(geno.comp.)<1) NA else {
    #fisher.test (  cbind ( compA1 , compB1 ) )
    (fisher.test (  geno.comp. )$p) }
```

```
} )
names(zeroVSsecond)<-colnames(GG_JL_All_Gy)

#0y vs 3y (No -log)
zeroVSthird <- sapply ( 1: ncol(GG_JL_All_Gy) , function(i) {
print(i)
  group.genotypes.A <- zero[,i]
  group.genotypes.B <- third[,i]
  genotype.freqs.A<-c(sum(group.genotypes.A %in% 0),
sum(group.genotypes.A %in% 1), sum(group.genotypes.A %in% 2))
  genotype.freqs.B<-c(sum(group.genotypes.B %in% 0),
sum(group.genotypes.B %in% 1), sum(group.genotypes.B %in% 2))
  geno.comp. <- cbind(genotype.freqs.A,genotype.freqs.B)
  if (sum(geno.comp.)<1) NA else {
    #fisher.test (  cbind ( compA1 , compB1 ) )
    (fisher.test (  geno.comp. )$p) }
} )
names(zeroVSthird)<-colnames(GG_JL_All_Gy)

## -Fisher's exact by age and sex- ##

#0yF vs 0yM (No -log)
zeroFVSzeroM <- sapply ( 1: ncol(GG_JL_All_Gy) , function(i) {
print(i)
  group.genotypes.A <- zeroF[,i]
  group.genotypes.B <- zeroM[,i]
  genotype.freqs.A<-c(sum(group.genotypes.A %in% 0),
sum(group.genotypes.A %in% 1), sum(group.genotypes.A %in% 2))
  genotype.freqs.B<-c(sum(group.genotypes.B %in% 0),
sum(group.genotypes.B %in% 1), sum(group.genotypes.B %in% 2))
   geno.comp. <- cbind(genotype.freqs.A,genotype.freqs.B)
  if (sum(geno.comp.)<1) NA else {
    #fisher.test (  cbind ( compA1 , compB1 ) )
    (fisher.test (  geno.comp. )$p) }
} )
names(zeroFVSzeroM)<-colnames(GG_JL_All_Gy)

#1yF vs 1yM (No -log)
firstFVSfirstM <- sapply ( 1: ncol(GG_JL_All_Gy) , function(i) {
print(i)
  group.genotypes.A <- firstF[,i]
  group.genotypes.B <- firstM[,i]
  genotype.freqs.A<-c(sum(group.genotypes.A %in% 0),
sum(group.genotypes.A %in% 1), sum(group.genotypes.A %in% 2))
  genotype.freqs.B<-c(sum(group.genotypes.B %in% 0),
sum(group.genotypes.B %in% 1), sum(group.genotypes.B %in% 2))
  geno.comp. <- cbind(genotype.freqs.A,genotype.freqs.B)
  if (sum(geno.comp.)<1) NA else {
    #fisher.test (  cbind ( compA1 , compB1 ) )
    (fisher.test (  geno.comp. )$p) }
} )
names(firstFVSfirstM)<-colnames(GG_JL_All_Gy)

#2yF vs 2yM (No -log)
secondFVSsecondM <- sapply ( 1: ncol(GG_JL_All_Gy) , function(i)
{ print(i)
  group.genotypes.A <- secondF[,i]
  group.genotypes.B <- secondM[,i]
  genotype.freqs.A<-c(sum(group.genotypes.A %in% 0),
sum(group.genotypes.A %in% 1), sum(group.genotypes.A %in% 2))
  genotype.freqs.B<-c(sum(group.genotypes.B %in% 0),
sum(group.genotypes.B %in% 1), sum(group.genotypes.B %in% 2))
  geno.comp. <- cbind(genotype.freqs.A,genotype.freqs.B)
```

```
  if (sum(geno.comp.)<1) NA else {
    #fisher.test (  cbind ( compA1 , compB1 ) )
    (fisher.test (  geno.comp. )$p) }
} )
names(secondFVSsecondM)<-colnames(GG_JL_All_Gy)

#3yF vs 3yM (No -log)
thirdFVSthirdM <- sapply ( 1: ncol(GG_JL_All_Gy) , function(i) {
print(i)
  group.genotypes.A <- thirdF[,i]
  group.genotypes.B <- thirdM[,i]
  genotype.freqs.A<-c(sum(group.genotypes.A %in% 0),
sum(group.genotypes.A %in% 1), sum(group.genotypes.A %in% 2))
  genotype.freqs.B<-c(sum(group.genotypes.B %in% 0),
sum(group.genotypes.B %in% 1), sum(group.genotypes.B %in% 2))
  geno.comp. <- cbind(genotype.freqs.A,genotype.freqs.B)
  if (sum(geno.comp.)<1) NA else {
    #fisher.test (  cbind ( compA1 , compB1 ) )
    (fisher.test (  geno.comp. )$p) }
} )
names(thirdFVSthirdM)<-colnames(GG_JL_All_Gy)

## -FOLLOWING AGE GROUPS MALES- ##

#0yM vs 1yM (No -log)
zeroMVSfirstM <- sapply ( 1: ncol(GG_JL_All_Gy) , function(i) {
print(i)
  group.genotypes.A <- zeroM[,i]
  group.genotypes.B <- firstM[,i]
  genotype.freqs.A<-c(sum(group.genotypes.A %in% 0),
sum(group.genotypes.A %in% 1), sum(group.genotypes.A %in% 2))
  genotype.freqs.B<-c(sum(group.genotypes.B %in% 0),
sum(group.genotypes.B %in% 1), sum(group.genotypes.B %in% 2))
  geno.comp. <- cbind(genotype.freqs.A,genotype.freqs.B)
  if (sum(geno.comp.)<1) NA else {
    #fisher.test (  cbind ( compA1 , compB1 ) )
    (fisher.test (  geno.comp. )$p) }
} )
names(zeroMVSfirstM)<-colnames(GG_JL_All_Gy)


#1yM vs 2yM (No -log)
firstMVSsecondM <- sapply ( 1: ncol(GG_JL_All_Gy) , function(i)
{ print(i)
  group.genotypes.A <- firstM[,i]
  group.genotypes.B <- secondM[,i]
  genotype.freqs.A<-c(sum(group.genotypes.A %in% 0),
sum(group.genotypes.A %in% 1), sum(group.genotypes.A %in% 2))
  genotype.freqs.B<-c(sum(group.genotypes.B %in% 0),
sum(group.genotypes.B %in% 1), sum(group.genotypes.B %in% 2))
  geno.comp. <- cbind(genotype.freqs.A,genotype.freqs.B)
  if (sum(geno.comp.)<1) NA else {
    #fisher.test (  cbind ( compA1 , compB1 ) )
    (fisher.test (  geno.comp. )$p) }
} )
names(firstMVSsecondM)<-colnames(GG_JL_All_Gy)

#2yM vs 3yM (No -log)
secondMVSthirdM <- sapply ( 1: ncol(GG_JL_All_Gy) , function(i)
{ print(i)
  group.genotypes.A <- secondM[,i]
  group.genotypes.B <- thirdM[,i]
  genotype.freqs.A<-c(sum(group.genotypes.A %in% 0),
```

```
sum(group.genotypes.A %in% 1), sum(group.genotypes.A %in% 2))
  genotype.freqs.B<-c(sum(group.genotypes.B %in% 0),
sum(group.genotypes.B %in% 1), sum(group.genotypes.B %in% 2))
  geno.comp. <- cbind(genotype.freqs.A,genotype.freqs.B)
  if (sum(geno.comp.)<1) NA else {
    #fisher.test (  cbind ( compA1 , compB1 ) )
    (fisher.test (  geno.comp. )$p) }
} )
names(secondMVSthirdM)<-colnames(GG_JL_All_Gy)

#0yM vs 3yM (No -log)
zeroMVSthirdM <- sapply ( 1: ncol(GG_JL_All_Gy) , function(i) {
print(i)
  group.genotypes.A <- zeroM[,i]
  group.genotypes.B <- thirdM[,i]
  genotype.freqs.A<-c(sum(group.genotypes.A %in% 0),
sum(group.genotypes.A %in% 1), sum(group.genotypes.A %in% 2))
  genotype.freqs.B<-c(sum(group.genotypes.B %in% 0),
sum(group.genotypes.B %in% 1), sum(group.genotypes.B %in% 2))
  geno.comp. <- cbind(genotype.freqs.A,genotype.freqs.B)
  if (sum(geno.comp.)<1) NA else {
    #fisher.test (  cbind ( compA1 , compB1 ) )
    (fisher.test (  geno.comp. )$p) }
} )
names(zeroMVSthirdM)<-colnames(GG_JL_All_Gy)

#0yM vs secondM (No -log) Important!
zeroMVSsecondM <- sapply ( 1: ncol(GG_JL_All_Gy) , Munction(i) { print(i)
 group.genotypes.A <- zeroM[,i]
 group.genotypes.B <- secondM[,i]
 genotype.Mreqs.A<-c(sum(group.genotypes.A %in% 0), sum(group.genotypes.A %in% 1),
sum(group.genotypes.A %in% 2))
 genotype.Mreqs.B<-c(sum(group.genotypes.B %in% 0), sum(group.genotypes.B %in% 1),
sum(group.genotypes.B %in% 2))
 geno.comp. <- cbind(genotype.Mreqs.A,genotype.Mreqs.B)
 iM (sum(geno.comp.)<1) NA else {
  #Misher.test (  cbind ( compA1 , compB1 ) )
  (Misher.test (  geno.comp. )$p) }
} )
names(zeroMVSsecondM)<-colnames(GG_JL_All_Gy)

## -FOLLOWING AGE GROUPS FEMALES- ##

#0yF vs 1yF (No -log)
zeroFVSfirstF <- sapply ( 1: ncol(GG_JL_All_Gy) , function(i) {
print(i)
  group.genotypes.A <- zeroF[,i]
  group.genotypes.B <- firstF[,i]
  genotype.freqs.A<-c(sum(group.genotypes.A %in% 0),
sum(group.genotypes.A %in% 1), sum(group.genotypes.A %in% 2))
  genotype.freqs.B<-c(sum(group.genotypes.B %in% 0),
sum(group.genotypes.B %in% 1), sum(group.genotypes.B %in% 2))
  geno.comp. <- cbind(genotype.freqs.A,genotype.freqs.B)
  if (sum(geno.comp.)<1) NA else {
    #fisher.test (  cbind ( compA1 , compB1 ) )
    (fisher.test (  geno.comp. )$p) }
} )
names(zeroFVSfirstF)<-colnames(GG_JL_All_Gy)

#1yF vs 2yF (No -log)
```

```
firstFVSsecondF <- sapply ( 1: ncol(GG_JL_All_Gy) , function(i)
{ print(i)
  group.genotypes.A <- firstF[,i]
  group.genotypes.B <- secondF[,i]
  genotype.freqs.A<-c(sum(group.genotypes.A %in% 0),
sum(group.genotypes.A %in% 1), sum(group.genotypes.A %in% 2))
  genotype.freqs.B<-c(sum(group.genotypes.B %in% 0),
sum(group.genotypes.B %in% 1), sum(group.genotypes.B %in% 2))
  geno.comp. <- cbind(genotype.freqs.A,genotype.freqs.B)
  if (sum(geno.comp.)<1) NA else {
    #fisher.test (  cbind ( compA1 , compB1 ) )
    (fisher.test (  geno.comp. )$p) }
} )
names(firstFVSsecondF)<-colnames(GG_JL_All_Gy)

#2yF vs 3yF (No -log)
secondFVSthirdF <- sapply ( 1: ncol(GG_JL_All_Gy) , function(i)
{ print(i)
  group.genotypes.A <- secondF[,i]
  group.genotypes.B <- thirdF[,i]
  genotype.freqs.A<-c(sum(group.genotypes.A %in% 0),
sum(group.genotypes.A %in% 1), sum(group.genotypes.A %in% 2))
  genotype.freqs.B<-c(sum(group.genotypes.B %in% 0),
sum(group.genotypes.B %in% 1), sum(group.genotypes.B %in% 2))
  geno.comp. <- cbind(genotype.freqs.A,genotype.freqs.B)
  if (sum(geno.comp.)<1) NA else {
    #fisher.test (  cbind ( compA1 , compB1 ) )
    (fisher.test (  geno.comp. )$p) }
} )
names(secondFVSthirdF)<-colnames(GG_JL_All_Gy)

#0yF vs 3yF (No -log)
zeroFVSthirdF <- sapply ( 1: ncol(GG_JL_All_Gy) , function(i) {
print(i)
  group.genotypes.A <- zeroF[,i]
  group.genotypes.B <- thirdF[,i]
  genotype.freqs.A<-c(sum(group.genotypes.A %in% 0),
sum(group.genotypes.A %in% 1), sum(group.genotypes.A %in% 2))
  genotype.freqs.B<-c(sum(group.genotypes.B %in% 0),
sum(group.genotypes.B %in% 1), sum(group.genotypes.B %in% 2))
  geno.comp. <- cbind(genotype.freqs.A,genotype.freqs.B)
  if (sum(geno.comp.)<1) NA else {
    #fisher.test (  cbind ( compA1 , compB1 ) )
    (fisher.test (  geno.comp. )$p) }
} )
names(zeroFVSthirdF)<-colnames(GG_JL_All_Gy)

#0yF vs secondF (No -log) Important!
zeroFVSsecondF <- sapply ( 1: ncol(GG_JL_All_Gy) , function(i) {
print(i)
  group.genotypes.A <- zeroF[,i]
  group.genotypes.B <- secondF[,i]
  genotype.freqs.A<-c(sum(group.genotypes.A %in% 0),
sum(group.genotypes.A %in% 1), sum(group.genotypes.A %in% 2))
  genotype.freqs.B<-c(sum(group.genotypes.B %in% 0),
sum(group.genotypes.B %in% 1), sum(group.genotypes.B %in% 2))
  geno.comp. <- cbind(genotype.freqs.A,genotype.freqs.B)
  if (sum(geno.comp.)<1) NA else {
    #fisher.test (  cbind ( compA1 , compB1 ) )
    (fisher.test (  geno.comp. )$p) }
} )
names(zeroFVSsecondF)<-colnames(GG_JL_All_Gy)
```

APPENDIX K: The R script for creating figures 6 and 7.

```
# Ordering the SNP containing loci by their genomic position
SET_FINAL2pos
index1 = order ( SET_FINAL2pos$pos ) # How big is the pos-
number, ascending. For example, the 109th line has the biggest
pos-number, thus the last in the vector and has value 109
index2 = order ( SET_FINAL2pos$chr [index1] )
index3 = index1 [index2] #combnined ordering index accounting
for both chm number an dposition in the chromosome
SET_FINAL2pos_ordered = SET_FINAL2pos [index3 , ]
SET_FINAL2pos_ordered
order_ind = order(match ( names(zeroVSparentM) ,
SET_FINAL2pos_ordered$id )) ## ordering plot. note text is not
ordered.
names.ordered.SNPs <- names(((-log(zeroVSparentM[order_ind]))))
SET_FINAL2_JAN = SET_FINAL2pos [ SET_FINAL2pos$id %in% sub("X" ,
"" , names(zeroVSadultM) ) , ]
order2 = order(SET_FINAL2_JAN$order  )
zeroVSadultM_ORD = zeroVSadultM [order2]
zeroVSadultF_ORD = zeroVSadultF [order2]

# Estimating the genetic inflation factor lambda
     # Males
med_null = qchisq   ( 0.5 , df = 2 ) ## q-value (X2 test
statistic) for p=0.5 and df=2 (you use gnotype)
med_obs = median ( qchisq(sort((zeroVSadultM_ORD)),2 ) )
GC_value = med_null / med_obs

     #Females
med_null_Jan_Female = median (qchisq(0.5,2)) ## q-value (X2 test
statistic) for p=0.5 and df=2 (you use gnotype)
med_obs_Jan_Female = median ( qchisq(sort((zeroVSadultF_ORD)),2
) )
GC_value_Jan_Female = med_null_Jan_Female / med_obs_Jan_Female

# ## confidances (MALE)
P_vals_conf_Chi = sapply (1:10000, function(x) {
  random.CHIstat =  rchisq( ncol(GG_JL_All_Gy) , 2, ncp = 0)
  random.CHIstat.P =  pchisq(random.CHIstat, 2, ncp = 0,
lower.tail = TRUE, log.p = FALSE)
  sort(random.CHIstat.P)
})

dim(P_vals_conf_Chi)
minP_Jan = -log( apply ( P_vals_conf_Chi ,   1 , function(x) {
sort(x) [500]  } ))
medP_Jan = -log( apply ( P_vals_conf_Chi ,   1 , function(x) {
sort(x) [5000] } ))
maxP_Jan = -log( apply ( P_vals_conf_Chi ,   1 , function(x) {
sort(x) [9500] } ))

GCmed_Jan = medP_Jan * GC_value_Jan
GCmin_Jan = minP_Jan * GC_value_Jan
GCmax_Jan = maxP_Jan * GC_value_Jan

# ## confidances (FEMALE)
minP_Jan_Female = -log( apply ( P_vals_conf_Chi ,   1 ,
function(x) { sort(x) [500]  } ))
medP_Jan_Female = -log( apply ( P_vals_conf_Chi ,   1 ,
function(x) { sort(x) [5000] } ))
```

```
maxP_Jan_Female = -log( apply ( P_vals_conf_Chi ,   1 ,
function(x) { sort(x) [9500] } ))
GCmed_Jan_Female = medP_Jan_Female * GC_value_Jan_Female
GCmin_Jan_Female = minP_Jan_Female * GC_value_Jan_Female
GCmax_Jan_Female = maxP_Jan_Female * GC_value_Jan_Female
########## PLOTS ############
#alpha treshold plot

par (mfrow=c(1,2))
  # MALE
plot ( -log(zeroVSadultM_ORD) , pch = 20 , xlab = "SNP index
orded by chromosomes" , ylab = "-log(P)" )
points ( x = IND_VIP , y = -log(zeroVSadultM_ORD) [ IND_VIP ]  ,
pch=20, col="red")
#abline ( h = -log(0.05 / length (zeroVSadultM_ORD)) * GC_value
, lty=2) # GC_value is calculated below. This is supre scrict
with bonferronni
abline ( h = -log(0.001) *  GC_value_Jan , lty=2)
abline ( h = -log(0.01) *  GC_value_Jan , lty=2)
text ( x = 35 , y= 18.5, labels = "alpha thershold = 0.001 x GC"
, cex= 0.7)
text ( x = 35 , y= 12.5, labels = "alpha thershold = 0.01 x GC"
, cex= 0.7)
text(x = IND_VIP, y = -log(zeroVSadultM_ORD) [ IND_VIP ], labels
= Simple_SNP_id, pos = 4, cex = 0.5) # UNFINISHED!
#legend("topleft", legend="a", box.lty=0, inset=.01,
text.font=2, cex=1.35)
  # FEMALE
plot ( -log(zeroVSadultF_ORD) , pch = 20 , xlab = "SNP index
orded by chromosomes" , ylab = "-log(P)" )
points ( x = IND_VIP , y = -log(zeroVSadultF_ORD) [ IND_VIP ]  ,
pch=20, col="red")
#abline ( h = -log(0.05 / length (zeroVSadultM_ORD)) * GC_value
, lty=2) # GC_value is calculated below. This is supre scrict
with bonferronni
#abline ( h = -log(0.001) *  GC_value_Jan_Female , lty=2)
abline ( h = -log(0.05) *  GC_value_Jan_Female , lty=2)
#text ( x = 35 , y= 17, labels = "alpha thershold = 0.01 x GC" ,
cex= 0.7)
text ( x = 35 , y= 3.58, labels = "alpha thershold = 0.05 x GC"
, cex= 0.7)
text(x = IND_VIP, y = -log(zeroVSadultF_ORD) [ IND_VIP ], labels
= Simple_SNP_id , pos = 4, cex = 0.5) # UNFINISHED!
#legend("topleft", legend=b)

# Q-Q PLOTS
par (mfrow=c(1,2))

  # Male

plot ( y = sort ( -log(zeroVSadultM_ORD) ) , x = rev(medP_Jan) ,
xlim = c(0,6) , type = "n" , ylab =  "observed -log(P)"  , xlab
=  "expected -log(P)" )
polygon ( y = c ( rev(GCmin_Jan) , GCmax_Jan   ) ,     x = c (
rev(medP_Jan) , medP_Jan ) , col = "light gray" , border = NA)
points ( y = sort ( -log(zeroVSadultM_ORD) ) , x = rev(medP_Jan)
,pch=20 , )
abline(0,1)#expcted
abline(0,GC_value_Jan , col="red")#inflated
text (x=1,y=17, labels = paste("GC =", format (GC_value_Jan,
digits=3)) )
points ( y = sort ( -log(zeroVSadultM_ORD) )[which(names(sort (
-log(zeroVSadultM_ORD) )) %in% IND_VIP_names)] , x =
```

```
rev(medP_Jan)[which(names(sort ( -log(zeroVSadultM_ORD) )) %in%
IND_VIP_names)] ,pch=20 , col="red")

# Female

plot ( y = sort ( -log(zeroVSadultF_ORD) ) , x =
rev(medP_Jan_Female) , xlim = c(0,6) , type = "n"  , ylab =
"observed -log(P)"  , xlab =  "expected -log(P)" )
polygon ( y = c ( rev(GCmin_Jan_Female) , GCmax_Jan_Female  ) ,
x = c ( rev(medP_Jan_Female) , medP_Jan_Female ) , col = "light
gray" , border = NA)
points ( y = sort ( -log(zeroVSadultF_ORD) ) , x =
rev(medP_Jan_Female) ,pch=20 , )
abline(0,1)#expcted
abline(0,GC_value_Jan_Female , col="red")#inflated
text (x=1,y=6, labels = paste("GC =", format
(GC_value_Jan_Female, digits=3)) )
points ( y = sort ( -log(zeroVSadultF_ORD) )[which(names(sort (
-log(zeroVSadultF_ORD) )) %in% IND_VIP_names)] , x =
rev(medP_Jan)[which(names(sort ( -log(zeroVSadultF_ORD) )) %in%
IND_VIP_names)] ,pch=20 , col="red")
```