OULU 2020
A 747

ACTA UNIVERSITATIS OULUENSIS

A

SCIENTIAE RERUM NATURALIUM

Jaakko Tyrmi

GENOMICS AND BIOINFORMATICS OF LOCAL ADAPTATION

STUDIES ON TWO NON-MODEL PLANTS AND A SOFTWARE FOR BIOINFORMATICS WORKFLOW MANAGEMENT

UNIVERSITY OF OULU GRADUATE SCHOOL;
UNIVERSITY OF OULU,
FACULTY OF SCIENCE;
BIOCENTER OULU

ACTA UNIVERSITATIS OULUENSIS
A Scientiae Rerum Naturalium 747

JAAKKO TYRMI

GENOMICS AND BIOINFORMATICS OF LOCAL ADAPTATION

Studies on two non-model plants and a software for bioinformatics workflow management

Academic dissertation to be presented with the assent of the Doctoral Training Committee of Health and Biosciences of the University of Oulu for public defence in the OP auditorium (L10), Linnanmaa, on 6 November 2020, at 4 p.m.

UNIVERSITY OF OULU, OULU 2020

Acta Univ. Oul. A 747, 2020

Supervised by
Docent Tanja Pyhäjärvi
Professor Outi Savolainen

Reviewed by
Assistant Professor Rosario Garcia-Gil
Assistant Professor Amanda De La Torre

Opponent
Assistant Professor Sam Yeaman

ISBN 978-952-62-2734-4 (Paperback)
ISBN 978-952-62-2735-1 (PDF)

ISSN 0355-3191 (Printed)
ISSN 1796-220X (Online)

Cover Design
Raimo Ahonen

PUNAMUSTA
TAMPERE 2020

Tyrmi, Jaakko, Genomics and bioinformatics of local adaptation. Studies on two non-model plants and a software for bioinformatics workflow management.
University of Oulu Graduate School; University of Oulu, Faculty of Science; Biocenter Oulu
Acta Univ. Oul. A 747, 2020
University of Oulu, P.O. Box 8000, FI-90014 University of Oulu, Finland

Abstract

It has been known for centuries that organisms from different geographic origins vary in fitness when transferred to another environment. Local adaptation – a situation where the local population fares best – is often observed. This phenomenon has a genetic basis but it is often not well understood. In this thesis, I examine the genetic basis of local adaptation using contemporary high-throughput sequencing and population genomic approaches and develop workflow management software to enable efficient computation of data from virtually any sequencing workflow. I study the population genetic features of local adaptation in Scots pine (*Pinus sylvestris*), a tree species with wide distribution range spanning from westernmost Europe to Eastern Eurasia. Similar questions are also addressed in a perennial herb, *Arabidopsis lyrata*.

In Scots pine, a targeted sequence capture was implemented, and divergence-based and landscape genomics methods uncovered several variants contributing to local adaptation. We also identified a very large inversion, which is potentially under selection. Whole genome sequencing of *Arabidopsis lyrata* uncovered demographic history and post-glacial colonization patterns in Northern Europe. Computation in both studies is largely automated and parallelized by STAPLER – the workflow management software produced in this thesis. The results highlight the benefits of allocating time in bioinformatics workflow design, the importance of developing novel methods to detect polygenic adaptation, and call for more frequent inclusion of analysis of structural variation in studies of local adaptation.

Keywords: bioinformatics, inversion, local adaptation, Pinus, Scots pine, selection, workflow

Tyrmi, Jaakko, Genomiikka ja bioinformatiikka. Nykyaikaisen lokaaliadaptaatiotutkimuksen kivijalat kahden kasvilajin ja laskennallisten työnkulkujen hallinnoinnin näkökulmasta.
Oulun yliopiston tutkijakoulu; Oulun yliopisto, Luonnontieteellinen tiedekunta; Biocenter Oulu
Acta Univ. Oul. A 747, 2020
Oulun yliopisto, PL 8000, 90014 Oulun yliopisto

Tiivistelmä

Jo vuosisatojen ajan on tiedetty organismien elinkelpoisuuden muuttuvan, mikäli niiden maantieteellistä sijaintia vaihdetaan. Usein havaittu ilmiö, lokaaliadaptaatio, tarkoittaa tilannetta, jossa lajin paikallisella populaatiolla on paras kelpoisuus muualta kotoisin oleviin populaatioihin nähden. Tällä ilmiöllä on geneettinen tausta, jota ei kuitenkaan tunneta erityisen hyvin. Tässä väitöskirjassa tutkin paikallisadaptaation geneettistä taustaa käyttäen nykyaikaisia sekvensointimenetelmiä. Kehitin tietokoneohjelman bioinformaattisten työnkulkujen automatisointiin ja parallelisointiin. Ensimmäinen tutkimuslaji, jonka populaatiogeneettisiä ominaisuuksia ja paikallisadaptaation genomisia merkkejä havainnoin on metsämänty (*Pinus sylvestris*). Sen levinneisyys ulottuu Länsi-Euroopasta aina Itä-Siperiaan saakka. Työssä etsitään vastauksia samankaltaisiin kysymyksiin tutkien myös toista lajia, monivuotista ruohovartista kasvia – idänpitkäpalkoa (*Arabidopsis lyrata*).

Metsämännyn genomia tutkittiin kohdennettua sekvensointimenetelmää käyttäen. Paikallisadaptaation merkkejä etsittiin maantieteellisten muuttujien ja geneettisten markkereiden välisiä korrelaatioita etsivien, sekä geneettiseen divergenssiin perustuvien tutkimusmenetelmien avulla. Lisäksi havaitsimme mahdollisesti luonnonvalinnan kohteena olevan suuren inversion. Idänpitkäpalkoa tutkittiin sekvensoimalla kasvien koko genomi. Tarkastelimme lajin demografista historiaa sekä jääkaudenjälkeisiä kolonisaatioreittejä pohjoisessa Euroopassa. Molempien töiden genomidatan bioinformaattinen analyysi suoritettiin supertietokoneympäristössä paralleelisti väitöstyössä kehittämääni STAPLER-työnkulkuohjelmaa käyttäen. Tämän väitöstyön tulokset korostavat bioinformaattisen työnkulun suunnittelun ja testauksen tärkeyttä sekä uusien menetelmien tarvetta polygeenisen adaptaation havaitsemiseksi. Samoin tulokset osoittavat, että rakenteellista geneettistä muuntelua on syytä tarkastella huolellisesti paikallisadaptaation perinnöllistä taustaa etsittäessä.

Asiasanat: bioinformatiikka, inversio, lokaaliadaptaatio, luonnonvalinta, metsämänty, työnkulku

To Eero, Sylvi and Kaarlo

Acknowledgements

First, I would like to thank my main supervisor Tanja Pyhäjärvi for all the support, encouragement and mentoring you have provided over the years. I would also like to thank my other supervisor Outi Savolainen for introducing me to population genetics and for guidance to scientific thinking.

I wish to thank my coauthors Tiina Mattila, Jaana Vuosku, Juan Acosta, Zhen Li, Lieven Sterck and Maria Cervera. Without your contribution this work would not have been possible. Special thanks to Soile Alatalo and Emmi Aikio for careful work in the laboratory. A warm thank you to Katri Kärkkäinen, Lumi Viljakainen and Andrew Eckert for all the support and good advice you gave me as my follow-up group members. I appreciate the valuable comments given by Amanda De la Torre and Rosario Garcia Gil after reviewing this thesis. My deepest thanks to Jaakko Lumme for saying, albeit jokingly, to a class of biology students roughly 12 years ago: “Learn how to program and everyone will want to hire you like crazy!”.

I’m grateful to the members of the Plant Genetics Group and the members of the Pyhäjärvi Lab for sharing your thoughts and expertise in our discussions, travels and events over the years. In particular, I would like to thank the Pine Troopers: Sonja Kujala, Komlan Avia, Yongfeng Zhou, Sandra Cervantes Arango, Dario Ojeda, Chedly Kastally and Robert Kesälahti. I would also like to acknowledge Komlan, Tiina and Iida Ronkainen for thorough testing and constructive feedback on STAPLER. I further thank Tiina for supervising my master’s thesis, for many interesting conversations and for peer support throughout the years. Special thanks to Tuomas Hämälä for invaluable advice in DNA-sequence analysis and for countless joyful discussions.

Suuri kiitos vaimolleni, vanhemmilleni ja ystävilleni. Kannustuksenne ja tukenne on ollut korvaamatonta työn loppuunsaattamisessa.

Oulu, September 2020 Jaakko Tyrmi

Publications

This thesis is based on the following publications, which are referred throughout the text by their Roman numerals:

I Tyrmi J.S., Vuosku J., Acosta J.J., Li Z., Sterk L., Cervera M.T., Savolainen O. & Pyhäjärvi T. (2020) Genomics of clinal local adaptation in *Pinus sylvestris* under continuous environmental and spatial genetic setting. G3: Genes | Genomes | Genetics. *https://doi.org/10.1534/g3.120.401285*

II Mattila T.M., Tyrmi J., Pyhäjärvi T. & Savolainen O. (2017) Genome-Wide Analysis of Colonization History and Concomitant Selection in *Arabidopsis lyrata*. Molecular Biology and Evolution. 34(10):2665–2677.

III Tyrmi, J. (2020) STAPLER: a simple tool for creating, managing and parallelizing common high-throughput sequencing workflows. Manuscript.

Table of contents

1 Introduction

For over 160 years, since Darwin first presented the concept of natural selection (Darwin, 1859), there has been an ongoing quest to better understand the of the ability of organisms to adapt to their local environmental conditions. This adaptation has genetic basis with different genetic variants affecting important traits, and subsequently the organism's probability to survive and produce offspring. The number of loci affecting each trait varies so that in rare cases there is only a single underlying locus, but most traits seem to be polygenic, meaning their variation is governed by several loci. The fate of a genetic variant (allele) is influenced by several factors, such as random sampling giving rise to fluctuation of allele frequency (genetic drift), influx of alleles from other populations (migration) and the strength of selection. There is a strong interest in discovering these variants (Rausher & Delph, 2015) in order to understand, not only how many loci are affecting each trait, but also to identify the types of variation ranging from simple single nucleotide polymorphism to more complex structural variants.

Tackling these questions often involves producing DNA-sequence data from many individuals, which together compose study populations. Next-generation sequencing approaches allow the generation of very large data sets, which are then shepherded through a bioinformatics workflow. This task can be aided by designing software to automate the management and parallelization of the often-complex workflows (Pfeifer, 2017). Large volumes of data prohibit large scale manual quality inspection and correction, thereby necessitating indirect and computational design approaches to reduce errors and minimize any potential bias in the results.

This thesis aims to firstly examine patterns of genetic diversity in *Pinus sylvestris*, a common tree species in northern Eurasia that is locally adapted to a wide range of environmental conditions found within its vast range (Pyhäjärvi, Kujala, & Savolainen, 2020). The goal is to examine the population structure and to uncover variants contributing to local adaptation. Secondly, the aim is to look into similar study questions, along with colonization history, of a perennial outcrossing herb *Arabidopsis lyrata*, with a strong emphasis on creating a high-quality workflow for the whole genome sequencing data set. Thirdly, the aim was to develop an easy-to-use workflow management software for automating and parallelizing bioinformatics workflows.

1.1 Local adaptation

Formally, local adaptation is a situation where populations have higher fitness at their home environment than any introduced populations (Kawecki & Ebert, 2004). Populations may become adapted to their local conditions if adaptive variation exists in the genome and if natural selection is strong enough relative to the genetic drift and gene flow. Classical work of Wright (1931) described how genetic drift, that is the random fluctuation of allele frequencies, can override the effects of natural selection. The effect of genetic drift is inversely proportional to the effective size of population, reducing the effect of selection in small populations. Equally classical work of Haldane (1930) examined the relationship between gene flow and selection. He showed that an allele favoured in a certain population would disappear if the rate of gene flow from non-local population would exceed the selection against it. In 1968 Levins described these effects by defining the concept of "environmental grain" (Levins, 1968), that shows how populations may be unable to adapt to finely grained environment in which spatial variation in the environment occurs at very short scale. Slatkin (1973) then made this idea more tangible by defining the equation 1,

$$l_c = l/\sqrt{s} \tag{1}$$

where *l* is dispersal distance, s is the strength of selection and l_c is the characteristic length, the minimum distance of environmental change to which allelic frequencies can respond to. Later, Slatkin (1978) started to expand this idea into traits controlled by multiple loci, but most theoretical work of polygenic adaptation has only been published in the recent decades.

To understand how local adaptation may arise and persist under spatially varying selection and gene flow, expectations for the underlying genetic architecture of polygenic traits must be defined. A well-known model presented by Fisher (Fisher, 1918) and refined by Orr (Orr, 1998) describes polygenic adaptation of a single population in a changing environment. The phenotype consists of a large number of continuous traits and fitness increases smoothly towards an optimum. A non-neutral mutation has a random effect on fitness, possibly increasing it if the genotype is brought closer to the optimum. The model predicts that adaptation will likely be due to few loci of large effect and many loci of small effect (Nicholas H. Barton & Keightley, 2002). The central idea of Fisher suggested that mutations of large effect were more likely to be harmful than small effect mutations, as they might overshoot the optimum even when the direction of mutation was correct, especially if the population is not far from the phenotypic optimum. However, Orr showed that some quite large effect loci are still expected to contribute, and to be fixed first (Orr, 1998).

Adaptation across environmental transects (for instance latitude, longitude or altitude) is a specific scenario, where means of traits related to local adaptation follow predictably varying local optima defined by abiotic and biotic conditions such as temperature, precipitation or levels of competition. Such a scenario is relevant for many species with wide distribution ranges. Several theoretical predictions about genetic architecture in these scenarios have been made. For instance, Barton (1999) proposed multiple models, of which some predict that a part of the loci would be fixed for one allele for much of the range and then have a rise in the frequency of the alternative allele, resulting in consecutive step clines. In other words, the habitat is divided into regions where different numbers of loci are nearly fixed for – and + alleles. It should be emphasized however, that only a subset of loci affecting a trait would be detectable via changes in allele frequency.

A body of literature by Latta (Latta, 1998, 2003), LeCorre and Kremer (2003, 2012; Kremer & Le Corre, 2012) describes how the contribution of covariance (linkage disequilibrium) between loci becomes more important relative to between population allele frequency differentiation, when the number of loci contributing to genetic architecture of a trait is large. As a result, as demonstrated by simulation experiments of Latta (1998), population differentiation of a locally adapted trait is poorly predicted from F_{ST} value of individual loci underlying the trait, particularly when there is high level of gene flow between populations. In such cases the between-population component in the trait value was mainly due to covariance of QTL loci allele frequencies between populations. F_{ST} values for QTL loci were higher than values for neutral markers only in cases where the strength of diversifying selection was extremely high. Also, some prediction of the sign of the covariance (i.e. positive or negative) could be made. For instance, more positive covariance results as the amount of gene flow increases. Strong diversifying selection tends to create more positive covariances (Latta, 1998). It should be noted however, that the sign of the covariance is relevant only if the directions of the allelic effects are known.

As a summary, the theoretical work of Orr, Barton and others show that some loci contributing to polygenic adaptation can cause a clear change in their allele frequency. Such signals may be detected using multiple population genetics approaches available, such as F_{ST} outlier tests (e.g. Foll & Gaggiotti, 2008) or landscape genetics methods (Rellstab, Gugerli, Eckert, Hancock, & Holderegger, 2015). On the other hand, the aforementioned models of Latta, LeCorre and Kremer do not necessarily predict such allele frequency differences.

The predictions of Orr were later revisited by Yeaman and Whitlock (2011), who showed how alleles of small effect contributing to local adaptation at the onset may be replaced by alleles of large effect. This may happen particularly when the traits experience stabilizing selection within populations, with gene flow occurring between populations. Large effect loci may consist of tightly linked alleles, as in the case of physical proximity or structural variation (Yeaman & Whitlock, 2011), as has also been suggested by some recent empirical findings (Samuk et al., 2017; Todesco et al., 2020). Inversions were recognized early on as possibly important form of variation, as it was hypothesized that they might form supergenes (Darlington & Mather 1949), defined by Dobzhansky (1970) as 'coadapted combinations of several or many genes locked in inverted sections of chromosomes and therefore inherited as single units'.

Kirkpatrick and Barton (2006) presented models to study how inversions containing beneficial variants may contribute to local adaptation in various scenarios. They show that if an inversion encompasses two or more locally adapted alleles, it has a fitness advantage over other haplotypes with fewer locally adapted alleles. A further contribution to the fitness advantage may occur if the inversion contains alleles with positive epistasis (Feldman, Otto, & Christiansen, 1997). Also, a particularly low deleterious mutation load within the inversion can lead to or contribute to a fitness advantage (Nei, Kojima, & Schaffer, 1967). If a fitness advantage arises, the inversion will rise in frequency within the geographic area where it provides a selective advantage until it reaches migration-selection balance. Locally advantageous mutations can also accumulate after initial spread (Faria, Johannesson, Butlin, & Westram, 2019). Inversions can create large areas of restricted recombination as they prevent proper chromatid pairing, although gene conversion and double crossovers may allow some genetic exchange between inverted and non-inverted haplotypes (Andoflatto, Depaulis, & Navarro, 2001). Empirical studies have shown that inversions contribute to local adaptation for instance in *Drosophila melanogaster* (Kapun & Flatt, 2018), sticklebacks (Jones et al., 2012), yellow monkeyflower (Gould, Chen, & Lowry, 2018), teosinte (Pyhäjärvi, Hufford, Mezmouk, & Ross-Ibarra, 2013), sunflowers (Todesco et al., 2020), humans (Puig, Villatoro, & Ca, 2015) and in many other species (Wellenreuther & Bernatchez, 2018).

Regardless of the type of polymorphism underlying adaptation, large proportion of contributing loci may remain undetected due to small changes in allele frequency, unless very large number of samples are genotyped and more advanced methods used (Berg & Coop, 2014). Furthermore, recent literature has indicated that the contribution of each loci can be transient, further complicating the efforts of detecting molecular signature of local adaptation (Barghi et al., 2019; Yeaman, 2015). In addition, a mutation beneficial throughout the species distribution range may produce a false signal of local adaptation at an early stage of its spread, when starting to spread from its geographic origin, as it has not yet spread widely (Booker, Yeaman, & Whitlock, 2019). It is also important to keep in mind the putative confounding effects of population history, as for instance an event like range expansion can also produce allele frequency clines at neutral loci (Excoffier & Ray, 2008).

1.2 *Pinus sylvestris* and *Arabidopsis lyrata* as study species

Pinus sylvestris L. is a conifer species native to Eurasia. It is a keystone species in many ecosystems within its huge distribution range spanning from mountains of southern Spain to taigas of eastern Siberia. Its distribution is mainly continuous, possibly with some level of gene flow throughout the range (Savolainen, Pyhäjärvi, & Knürr, 2007), but isolated populations exist outside the main range. The current census population size has been roughly estimated to be in the order of several hundred billion (Pyhäjärvi et al., 2020). Previous investigations have shown that repeated glaciation cycles have caused fluctuation in population sizes and impacted the population structure observed at mitochondrial level (Cheddadi et al., 2006; Naydenov, Senneville, Beaulieu, Tremblay, & Bousquet, 2007; Savolainen & Pyhäjärvi, 2007). However, investigations of the nuclear genome have shown only minimal population structure within the main range, although more distinct patterns can be detected in the isolated populations (Karhu et al., 1996; Dvornyk et al., 2002; Savolainen & Pyhäjärvi, 2007; Kujala & Savolainen, 2012).

Decades of forestry research have shown that the phenotypic signal of local adaptation is prominent in *P. sylvestris* (Savolainen et al., 2007; Pyhäjärvi et al., 2020), although the evidence has largely come indirectly from provenance trials rather than reciprocal transplant experiments, which are considered to be the most reliable method for observing local adaptation (Blanquart, Kaltz, Nuismer, & Gandon, 2013). Considerable phenotypic variation can be seen in traits such as phenology (Beuker, 1994; Karhu et al., 1996; Mikola, 1982), cold tolerance (Aho, 1994; Eiche, 1966; Hurme, Repo, Savolainen, & Pääkkönen, 1997; Hurme, Sillanpää, Arjas, Repo, & Savolainen, 2000) and many others (as reviewed in

Pyhäjärvi et al., 2020), with much of this variation being clinal along latitude. Genetic basis of the adaptation has remained elusive due to limited genomic resources available for *P. sylvestris* and possibly due to methodological difficulties of detecting polygenic signals of adaptation.

The genome size of *P. sylvestris* is estimated to be 23.6 Gbp (Zonneveld, 2012). Pines in general have high repetitive sequence content and over 50,000 genes (Wegrzyn et al., 2014; Stevens et al., 2016; Ojeda et al., 2019). Much of the repetitive sequence seems to be due to transposable elements, duplicated genes, pseudogenes and ancient gene duplications (Z. Li et al., 2015; Pavy et al., 2017; Zimin et al., 2014). Due to these factors, significant proportion of the conifer genomes are paralogous. There is no reference genome available for *P. sylvestris*, but reference genome of a closely related *Pinus taeda* (Neale et al., 2014) can be used for e.g. read mapping. This combination of huge repetitive genomes and lacking genomic resources makes genome-wide analysis a challenging task on *P. sylvestris*.

A. lyrata is a perennial outcrossing herb and a close relative of the model species *Arabidopsis thaliana*. It has a wide distribution throughout Eurasia and North America, but the distribution is patchy with highly isolated populations leading to strong population structure (Muller, Leppälä, & Savolainen, 2008; Pyhäjärvi, Aalto, & Savolainen, 2012; Ross-Ibarra et al., 2008). Local adaptation has been demonstrated among populations throughout its distribution (Hämälä, Mattila, & Savolainen, 2018; Leinonen, Remington, & Savolainen, 2011; Leinonen et al., 2009) even within short geographical scale under gene flow (Hämälä & Savolainen, 2019).

A. lyrata has excellent genomic resources, including a high-quality reference genome (Hu et al., 2011). The genome size of *A. lyrata* is comparatively small with a length of 207 Mbp and over 30,000 predicted genes. Furthermore, much of the vast resources available for the model species *A. thaliana* can also be utilized in *A. lyrata* studies as the species are closely related.

1.3 A brief history of DNA sequencing

Development of Sanger's chain-termination sequencing in the 70s (Sanger, Nicklen, & Coulson, 1977) was a major breakthrough in genetics, as it enabled investigators to produce high quality DNA sequence with relative ease. It thus became the dominant sequencing approach for the following decades. Over the years several advances, such as PCR amplification (Saiki et al., 1988, 1985) and enhanced automation of the sequencing instruments (Smith et al., 1986) increased the throughput of the method. Even though the length of the produced reads was relatively long 800 base pairs with excellent accuracy, only few Mbp of sequence could be produced with Sanger-sequencing machines per day (Kircher & Kelso, 2010), prohibiting genome-wide studies for most studies/species.

Multiple new second generation sequencing methods emerged and became commercially available in the early 2000s, namely the 454-sequencing (Margulies et al., 2005), SOLiD-sequencing (Shendure et al., 2005), and Solexa sequencing, which was later bought by Illumina (Bentley et al., 2008) and proceeded to become the most popular choice of that generation of methods. The throughput of the Illumina sequencing machines is up to 1600 Gbp of sequence per day using a NovaSeq 6000 instrument – several orders of magnitude more compared to Sanger sequencing. However, the read lengths of the highest throughput instruments are no more than 100-250 base pairs, complicating reference genome building and resequencing efforts, particularly in organisms whose genome contain high amount of repetitive sequence.

The most recent generation of sequencing methods not only provide higher throughput, but more importantly, aim to overcome the inconvenience and limitations imposed by the short reads. Pacbio sequencing (Eid et al., 2009) can produce reads with lengths up to 150 kbp and Oxford NanoPore sequencing (Clarke et al., 2009) produces reads up to two Mbp (Lu, Giordano, & Ning, 2016). Additional examples are BioNano (BioNano Genomics, San Diego, California) and Hi-C (Lieberman-Aiden et al., 2009) long range mapping methods, that provide the ability for improving the difficult-to-assemble plant reference genomes, although all of the newer methods also come with distinct error profiles and other caveats, which must be appropriately accounted for (Jung, Winefield, Bombarely, Prentis, & Waterhouse, 2019).

1.4 Role of bioinformatics analysis in population genomics

After the Sanger sequencing was introduced in 1977 (Sanger et al., 1977), biologists could during the following decades examine DNA sequence variation much more efficiently than before. Still the sizes of the datasets from these sequencing machines were manageable so that they could be entirely visually inspected for any technical issues. Computer software also improved so that much of the subsequent analysis could often be performed within dedicated software such as BIOSYS-1 (Swofford & Selander, 1981), MEGA (Kumar, Tamura, & Nei, 1994) or GENEPOP (Raymond, 1995).

As discussed above, the use of high-throughput sequencing data has become ubiquitous in the field of genetics, providing a more comprehensive genome-wide perspective for investigators. These massive datasets expose bioscientists to new computational challenges they may not have traditionally had to face. Firstly, the process of identifying genetic variation from raw sequencing data requires a multitude of processing steps, but errors and in some cases severe bias may be introduced to the dataset if proper care is not taken during the analysis (Pfeifer, 2017). Secondly, the processing steps are computationally intensive, especially in species with large genomes, requiring the use of supercomputing platforms. Using these requires at least rudimentary understanding of computer architecture and UNIX operating systems. This task can be made more manageable and user-friendly with appropriate workflow management software.

1.4.1 Bioinformatics workflow in resequencing studies

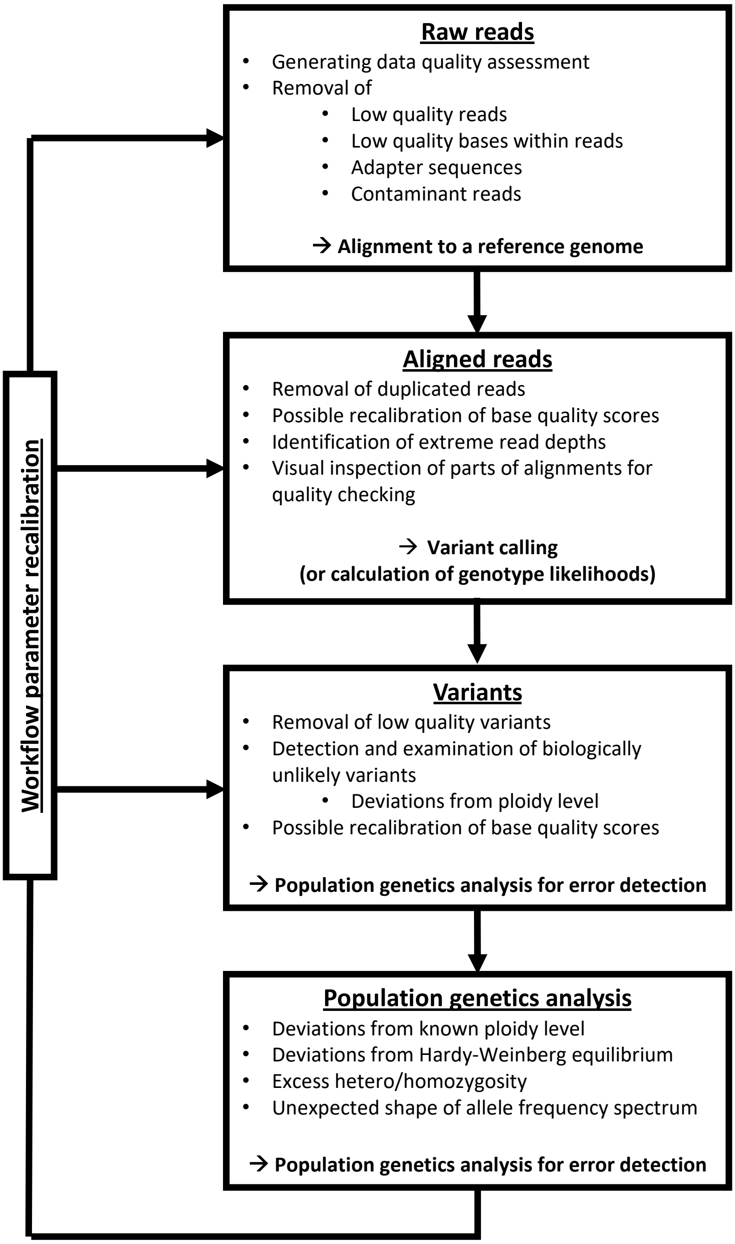
A workflow for uncovering genetic variants from raw sequence data usually includes multiple steps illustrated in Figure 1, although population genetics approaches can be used as a further indicator of putative technical problems as shown in the figure and discussed at the end of this chapter. First, the raw reads are processed by removing any technical sequences, like adapters used in attaching reads to sequencing platform, removing low quality reads or parts of reads with low quality. At this point reads should also be analysed for any systematic biases in the sequence contents, which may point to issues in preparing the sequencing library. However, particularly in the commonly used Illumina sequencing (Bentley et al., 2008), substitutions are the primary error type (Dohm, Lottaz, Borodina, & Himmelbauer, 2008), which cannot be reliably identified and removed from raw reads.

Next, the processed reads are aligned to a reference genome. Common steps at this point are refining base quality scores and removing read duplicates. At this stage several issues may arise. In positions where an insertion or a deletion has occurred the multiple equally good alignments of reads may exist. In this case the aligner software is often capable of choosing one alternative and then aligning all reads of the sample the same way. As the alignment is commonly done independently to each sample, the aligner may choose a different alignment in different samples resulting in spurious polymorphic variant calls. This issue is remedied in most variant calling software, as they contain features for realigning indel areas consistently in all samples before genotype calling (Garrison & Marth, 2012; McKenna et al., 2010). A second, and often the more serious issue, is an incorrect alignment of paralogous sequences. This is caused by the incompleteness of the reference sequences, as repetitive areas of the genome are technically difficult to separate during sequence assembly. This issue does not concern only areas of short repeats, but also coding areas if there has been whole genome or single gene duplications in recent history (Ojeda et al., 2019). If multiple paralogous sequences have been clustered together in the assembly, the reads originating from distinct parts of the genome will be aligned into this single location, causing spurious, often heterozygous, variant calls. Several approaches exist for removing this bias from variant calls. As reads originating from multiple parts of the genome align to a single location, a high read depth may be observed in such areas.

The variant call step is often the most time consuming as, unlike earlier steps, this is usually done jointly for all samples. A central dilemma in resequencing studies is the sequencing depth, as funds available for sequencing are finite, but high read depths result in more confident variant calls – especially in diploid samples where the calling of heterozygote requires at least a depth of 20 (Bentley et al., 2008). Very low read depths are problematic also for haploid samples, as the potential substitution errors often seen only in single reads may then have higher chance of causing spurious variant calls. A viable solution for analysing low depth data is to not use discrete variant calls in subsequent population genetics analysis, but utilize genotype likelihoods which incorporate this statistical uncertainty, as implemented for instance by Korneliussen et al. (Korneliussen, Moltke, Albrechtsen, & Nielsen, 2013).

The main difficulty in identifying many of the aforementioned issues in resequencing data processing is the vast amount of data that prohibits close examination of the whole raw data. Instead, investigators should carefully check the various metrics and summary information many of the tools produce for any anomalies. Visualization of the reads, alignments and variant calls is, although not strictly speaking mandatory, a crucial part of the workflow. Vigilance is also required in interpreting the results of any downstream analysis, as some issues, like spurious variants caused by paralogous mapping, may slip through the various quality metrics unnoticed. However, it is possible to detect the presence of spurious variant calls, for instance, as a surprisingly high level of nucleotide diversity or surprisingly large deviations from Hardy-Weinberg equilibrium (HWE)

(McKinney, Waples, Seeb, & Seeb, 2017). However, the fact that paralog alignment issue tends to occur in most or even in all samples in the same area, it can affect the shape of allele frequency spectrum so that intermediate frequency alleles are over-represented. A careful HWE-based filtering can be applied, for example by removing variants where all samples have heterozygous genotype – an extremely unlikely event to occur by chance in an outcrossing population. Also, a dedicated method, such as HDplot (McKinney et al., 2017) can be used. Alternatively, various filters can be applied to the bioinformatics workflow and their effect can be observed in the deviation of allele frequency spectrum. Filtering variants using HWE based thresholds maybe reasonable in some situations, but if care is not taken, in most population genomics applications of resequencing data such filtering can mask areas of biological interest, as natural causes like overdominance may produce similar patterns at individual loci. Even worse, applying filters that affect the allele frequency spectrum shape heavily can easily introduce further bias into the dataset, affecting for instance inference of demographic history analyses. In the case of haploid samples, filtering may be more straightforward as any presence of heterozygous variants suggests that incorrect variants are present within the region.



Raw reads
Generating data quality assessment
Removal of
Low quality reads
Low quality bases within reads
Adapter sequences
Contaminant reads
→ Alignment to a reference genome
Aligned reads
Removal of duplicated reads
Possible recalibration of base quality scores
Identification of extreme read depths
Visual inspection of parts of alignments for quality checking
→ Variant calling (or calculation of genotype likelihoods)
Variants
Removal of low quality variants
Detection and examination of biologically unlikely variants
Deviations from ploidy level
Possible recalibration of base quality scores
→ Population genetics analysis for error detection
Population genetics analysis
Deviations from known ploidy level
Deviations from Hardy-Weinberg equilibrium
Excess hetero/homozygosity
Unexpected shape of allele frequency spectrum
→ Population genetics analysis for error detection
Workflow parameter recalibration

Fig. 1. Steps of bioinformatics workflow in a population genomics study.

1.4.2 Key concepts of workflow management frameworks

Implementing the workflow from raw sequence reads to variant calling requires not only understating of the significance of the individual steps and possible caveats but also technical expertise and time. The computational runtime from start to finish of a workflow may be quite significant, but often the most time-consuming stages are prototyping different workflows by experimenting with different software and parameter combinations and attempts for sufficient parallelization and troubleshooting.

There are several ways to define and execute a workflow. For instance, an intuitive and possibly the most common way to create workflow is to write a shell script. These often simple text files may just list the various commands of the workflow or incorporate some basic programming in the form of UNIX shell scripting. Running the script will then execute the workflow. However, advanced workflow frameworks have been developed, and they often include many desirable qualities, such as modularity, optimal use of computing resources, reporting of finished runs and the ability to rerun failed parts of the workflow.

Modularity implies that a varying number and type of input files can be accommodated into the workflow, and more importantly, that the tools incorporated into the workflow can be freely selected and interchanged. The former feature is ubiquitous in workflow frameworks, but the latter is often absent, resulting in hard-coded workflows. These hard-coded tools are tailored by the authors to accomplish specific tasks (Fisch et al., 2015; Monat et al., 2015; Bourgey et al., 2019) with predefined set of tools and parameters. Such frameworks can be very useful in very specific scenarios, or when working with model species for which the hard-coded approach has been demonstrated to work. They may, however, be inadequate for non-model species datasets requiring more customized approaches and may quickly become obsolete when new analysis software is published but the relevant parts of the built-in workflow cannot be updated.

Allowing the user to choose the software for each step of the workflow and to define parameters freely provides flexibility, but the development of such workflow frameworks is non-trivial as there is more user input to inspect, and more importantly, they may be more difficult to use as the user has a plethora of options and parameters to digest. Furthermore, the user must understand the basics of supercomputing platform architecture. Such existing frameworks can be divided to command line systems aimed at technically adept computational investigators (Köster & Rahmann, 2012; Sadedin, Pope, & Oshlack, 2012), and easy-to-use systems with graphical user interface (Oinn et al., 2004; Goecks et al., 2010). The former type of framework provides great flexibility but requires programming expertise and the latter, while easier to use, may not be permitted on many supercomputer platforms, unless explicitly made available by the computing service provider.

The large high-throughput sequencing datasets are often processed on supercomputers or superclusters. These platforms are almost always run by dedicated staff and access to these resources are commonly only provided via a workload manager software such as SLURM, SGE or TORQUE. The workload managers penalize users who spend a disproportionate amount of computing time or memory by giving lower priorities to their jobs. As memory and processor core count requirement vary drastically within most bioinformatic pipelines, frameworks must interact with the workload managers and ensure efficient use of computational resources by automatically or by request splitting the workflows into parts, in which computational requirements in terms of memory or thread use are similar. This ensures that the resources required for the most computationally intensive part of a workflow are not reserved throughout the whole workflow run.

Workflows often consist of a large number of steps (Pfeifer, 2017) and large datasets, and so runtime and computational cost of executing the workflow may be considerable. It is therefore preferable that workflow frameworks perform a sanity check of the designed workflow, ensuring that an output of a given step is an eligible input for the next step before running. In the case of failure during runtime, it is important to have log reports with detailed information on at what stage the process failed and on putative underlying reasons. Also, a key feature in workflow manager is the ability to rerun only the failed part of the pipeline as rerunning the whole pipeline will waste computational resources and time.

Lastly, if the necessary features are in place, ease of use can significantly improve the likelihood of users adopting new software (Davis, Bagozzi, & Warshaw, 1989). This probably applies in the bioinformatics field particularly to workflow management software, as its adoption is in most cases entirely optional. Especially non-computational investigators – who in fact might benefit from such software the most – may not find it worthwhile to spend the time learning complicated software. Ease-of-use can be improved by software design and also by well written documentation and tutorials.

1.5 Aims of the study

This thesis is interdisciplinary, consisting of biological topics of population genomics and local adaptation, bioinformatics topics of resequencing and targeted sequence capture data processing and a software design topic of workflow management software development.

In the first part (I), a targeted sequence capture of *P. sylvestris* populations from wide geographic area, in two parallel clines, was conducted to study patterns of genetic diversity to answer the following questions: 1) Do the uncovered levels of genome-wide genetic diversity or LD patterns indicate the impact of local selection? 2) Have dispersal patterns (wind pollination) and the continuous distribution, common in tree species, resulted in spatially continuous isolation-by-distance pattern across the genome and distribution range? 3) Do we see genomic signatures of local adaptation in the form of allele frequency clines, F_{ST} outliers or differentiation of structural variation in *P. sylvestris*?

In the second part (II), a whole genome resequencing study of multiple *A. lyrata* populations was performed to study demographic history and local adaptation. In this study I concentrated on development of bioinformatics workflow to ensure a high-quality dataset by experimenting with several different workflows.

In the third part (III), the goal was to develop a workflow manager software aimed particularly for managing bioinformatic workflows in resequencing studies with main priorities ease-of-use, modularity, and efficient parallelization options. The aim was also to exploit this software in parts I and II to first enable prototyping of alternative bioinformatic pipelines and then allow the analysis of the whole dataset with the best discovered workflow.

2 Materials and methods

The workflow and materials are described only briefly in this section. The original papers provide more details, with the exception of study II, where some more details of the workflow design are explained here.

2.1 Study populations, samples and sequencing

Samples for study I originated from 12 *P. sylvestris* populations, of which the five easternmost ones originated from Russia, four from Finland, one from Latvia, one from Poland and one from Spain. The sampling was designed to maximize odds for detecting allele frequency clines by choosing sampled populations along two latitudinal gradients, one formed by the Russian samples and the other formed by the European samples, with the exception of the single population from Spain. The Spanish population is located within the Sierra Nevada mountains and is isolated from the continuous main distribution of *P. sylvestris*. From each population 10 samples were obtained. Haploid DNA material of megagametophyte tissue found within seeds was used in the sequencing. The use of haploid material is useful in population genetics study as the haplotypes are readily available without the need for phasing the data and also because many technical issues, such as indel and paralog alignment issues, manifest as heterozygous variants not expected in haploid data, and can therefore be removed.

Targeted sequence capture for coding areas known *P. sylvestris* transcripts was designed in cooperation with MycroArray MYbaits (Ann Arbor, MI). The performance of the baits was then evaluated in pilot sequence capture experiments, and as a result 60,000 baits were selected for the final design. Captured DNA fragments of the 120 samples were then sequenced with Illumina HiSeq 2500 instrument at Institute of Molecular Medicine Finland (FIMM).

In study II *A. lyrata* samples originated from five populations located in USA, Great Britain, Austria, Germany, Norway and Sweden. The north American samples are from subspecies *lyrata* and Eurasian samples from subspecies *petraea*. Whole genome resequencing data was obtained for 6 individuals from each population using Illumina HiSeq2000 instrument at FIMM, except in the British population where two samples were available.

2.2 Bioinformatics workflow

2.2.1 *Development of workflow manager software*

In study III a workflow manager software “STAPLER” was developed with Python 2 programming language to easily create, prototype and parallelize bioinformatics pipelines. In brief, the user creates a workflow by defining input files, software steps for processing the data and further details related to parallelization. Based on this information STAPLER will apply the workflow for each input file and ensures that the workflow is run in parallel. A directory tree is created for intermediate and final output files.

STAPLER was developed by applying an object-oriented programming (OOP) approach, which greatly aids in modelling of concepts and real-world object into a computer program. In brief, OOP code is not organized into functions and logic as in procedural programming, but rather into “classes”. A class may provide a definition for a real world or an abstract concept and contains any related data and behaviour. Specific instances of these classes, “objects”, can then be created, that contain the data of a specific case. In STAPLER, two kind of concepts were modelled as classes: bioinformatics tools and directory trees.

Each supported bioinformatics tool was modelled as a “tool” class defining what kind of parameters and input files are required and what type of output files will be produced. When an object is generated of such class, it contains the parameters the user defined in the input file for the given step, and also the paths to input and output files. In addition, they have the ability (i.e. *behaviour* in OOP terminology) to notify the user if certain parameters are missing or are not recognized by the tool in question.

As bioinformatic pipelines consist of multiple steps where the output of first step is the input of second step and so forth, it was necessary for STAPLER to predict the file structure of a workflow before it is actually created to check that valid inputs would exist for each program. Thereby directory trees of workflows were also modelled within STAPLER by directory and file classes. Objects of directory classes could contain objects of file class or other directory objects. Each directory object contains information about their directory path and possible contents, with the ability of adding or removing elements and telling whether a certain tool object has already taken the file as input to avoid conflicts of two commands taking the same file as input.

STAPLER is freely available at https://github.com/tyrmi/STAPLER. Bioinformatic workflows of studies I and II were built, parallelized and run using STAPLER.

2.2.2 Data analysis workflow

In study I the raw reads were aligned to the reference genome of the closely related *P. taeda* (v.1.01) (Neale et al., 2014), as there is no reference genome available for *P. sylvestris*. The raw read technical quality was analysed with FastQC and Fastx tools, after which they were aligned to the reference genome using bowtie2 version 1.1.1 (Langmead & Salzberg, 2012). The alignment files were then converted to BAM format, sorted, deduplicated and indexed with samtools (Li et al., 2009) and Picard-toolkit (http://broadinstitute.github.io/picard/). Alignment files were visualized with samtools tview, which uncovered severe alignment issues throughout the targeted regions, because paralogous sequences were captured by the baits and subsequently aligned such sequence to incorrect places. Such areas were identified by first performing a specific variant call with freebayes (Garrison & Marth, 2012) to uncover heterozygous variant calls, suggesting technical issues in haploid dataset. Variant call was then redone for settings suitable for haploid data, but with the heterozygous areas omitted. During the workflow ten samples were removed due to low technical quality and one sample was removed due to the same tree being accidentally sampled twice.

In study II the raw read quality was first checked using FastQC and Fastx toolkits. Reads were trimmed using trimmomatic (Bolger, Lohse, & Usadel, 2014) and mapped to *A. lyrata* reference genome Ensembl plant version 1.0.29 with bwa-mem (Li & Durbin, 2010). The alignment files were then converted to BAM format, sorted, deduplicated and indexed with samtools (Li et al., 2009) and Picard-toolkit (http://broadinstitute.github.io/picard/). For clipping overlap of forward and reverse reads BamUtil (http://genomes.sph.umich.edu/wiki/BamUtil) was used. Indels were realigned with Genome Analysis Toolkit (McKenna et al., 2010). No variant calling was done, instead all analysis relied on genotype likelihoods produced with Genome Analysis Toolkit or with ANGSD (Thorfinn Sand Korneliussen, Albrechtsen, & Nielsen, 2014). An excess of heterozygotes was visually observed in the intermediate frequency sites of the allele frequency spectrum, suggesting that the dataset was suffering from paralogous sequence alignment. Repetitive areas, and areas smaller than 200 bp between repetitive areas, were thereby masked to avoid potential mapping errors. The read depth for the whole alignments for each individual was calculated, and areas where read depths exceeded 3 times median absolute deviation (MAD) for 10 consecutive sites were also masked. After filtering the excess of heterozygotes could still be observed in the allele frequency spectrum. Therefore, a variant call was performed with freebayes to detect and remove positions where all samples have heterozygote genotype. There was also Sanger sequence available from two samples used in this study for 37 genes. As Sanger sequence has lower error rate by an order of magnitude compared to Illumina sequencing, and as the alignments of the Sanger reads were carefully manually inspected, we were able to use the Sanger data as a ground truth guide in designing the workflow.

2.3 Estimating genetic diversity and population structure

In study I, for characterizing genetic diversity of the populations, pairwise nucleotide diversity (Nei & Li, 1979) and Tajima's D (Tajima, 1989) were calculated with the program ∂a∂i (Gutenkunst, Hernandez, Williamson, & Bustamante, 2009). Pairwise F_{ST} values (Hudson, Slatkin, & Maddison, 1992) were calculated for each population pair. Population structure was further studied by using principal component analysis (McVean, 2009) with the R package prcomp and STRUCTURE (Pritchard, Stephens, & Donnelly, 2000). Further analysis of the populations structure was performed with the method conStruct (Bradburd, Coop, & Ralph, 2018), which allows accounting for the presence of isolation-by-distance. Linkage disequilibrium patterns were estimated with allelic frequencies correlation coefficient r^2, which was calculated between all variants within the same scaffold over all populations.

In study II population structure was explored using similar methods as in study I, but the analyses were conducted in sliding windows within the genotype call free ANGSD software (Korneliussen et al., 2013), where applicable. However, in this study ADMIXTURE (Alexander, Novembre, & Lange, 2009) and NGSadmix (Skotte, Korneliussen, & Albrechtsen, 2013) were used instead of STUCTURE and conStruct tools. Additionally, the Stairway plot method (Liu & Fu, 2015) was used to infer the historical sizes of individual populations. SFS based composite likelihood demographic modelling method, as implemented in fastsimcoal2 (Excoffier, Dupanloup, Huerta-S??nchez, Sousa, & Foll, 2013), was used to infer the population split history.

2.4 Uncovering the genetic basis of local adaptation

In study I, latitudinal allele frequency clines were searched for with the regression model in R package lme4. The landscape genomics approach bayenv (Coop, Witonsky, Di Rienzo, & Pritchard, 2010) was applied, but the results were discarded as unreliable due to inconsistencies between separate runs. It is possible that due to very limited population structure within the main distribution of *P. sylvestris,* the population matrix could not be reliably estimated leading to overcorrection and inconsistent behaviour between runs, although this could not be confirmed. The F_{ST} outlier method Bayescan (Foll & Gaggiotti, 2008) was used to detect putative loci responsive for local adaptation, along with Pcadapt (Luu, Bazin, & Blum, 2017), which corrects for the confounding effects of population structure via PCA.

The Bayescan and PCAdapt results uncovered a large haplotype structure within the *P. sylvestris* genome, with SNPs in full linkage disequilibrium across multiple scaffolds. The likelihood for such event occurring by chance was tested with a permutation test, and a rough position in the genome and a size estimate for the structure was estimated by placing the markers on genetic map published by Westbrook et al. (2015).

In study II, patterns of selection in the genome were searched for using population branching statistics (Yi et al., 2010) in sliding windows.

3 Results and discussion

3.1 STAPLER software speeds up prototyping and execution of bioinformatic workflows

The python program STAPLER, developed in study III and used in generating and parallelizing bioinformatics pipelines in studies I and II, is a fully featured workflow management software. It provides utilities to validate that workflow run has finished successfully, to rerun failed parts of the workflow and to compress intermediate files to save storage space. Multiple workflows can be quickly prototyped by copy-pasting and modifying the parameters. To make the software accessible on most platforms, it was written with Python 2 programming language. Python is an interpreted language, meaning that the program does not require compiling the code into machine-language instructions before use, but requires an interpreter software for running. Python 2 interpreter is preinstalled on almost all Linux, MacOS and other UNIX or UNIX related operating systems.

Input for STAPLER is a workflow text file that describes the directory path to input data, steps of the workflow as command lines and possible parameters resource manager when parallelizing workflows in supercomputing environments. Command line steps for each workflow can be defined by the user by replacing the input and output paths by keywords “$INPUT” and “$OUTPUT” which are then automatically replaced by actual file paths by STAPLER. STAPLER also natively supports select commonly used bioinformatics software. If natively supported software is incorporated into a workflow, STAPLER will also check that all parameters required to run the software are present. The workflow is completely modular, giving the user the freedom to select the order and content of pipelines without restrictions.

Workflows can be parallelized either as UNIX background processes or as batch jobs in supercomputing environments that use SLURM, LSF or TORQUE as workflow managers. If the parallelization is done as UNIX background job, STAPLER automatically detects the number of computer cores available and will split task accordingly. In the case of utilizing a resource manager the user will have to manually define how many cores will be utilized by the workflow.

The feature set is similar to some other workflow managers, with probably TOGGLE (Monat et al., 2015) being the closest in features, although not quite matching the ease-of-use and modularity of STAPLER. Other alternatives include bpipe (Sadedin et al., 2012) and Snakemake (Köster & Rahmann, 2012), which provide greater flexibility and configurability compared to STAPLER but require programming expertise to use. Taverna (Oinn et al., 2004) and Galaxy (Goecks et al., 2010) provide easy-to-use graphical user interface (GUI) and some powerful features such as branching of workflows, but may not be easy to install on all platforms and many users may prefer the use of command line over GUI.

STAPLER was well suited for analyzing the data sets in both *P. sylvestris* and *A. lyrata* studies despite the differences in the datasets, as the type and properties of the processed data sets makes little difference from the perspective of using STAPLER. In addition to the aforementioned data sets, STAPLER has also been used in creating analysis workflow for population genetic analysis of *Capsella grandiflora* (Mattila et al., 2019). Outside population genomics studies, the software is currently applied in automating and parallelizing GWAS and meta-analysis work unraveling genetic architecture governing human circulatory lipidome, and identifying risk alleles for medical conditions, such as pre-eclampsia, uterine fibroids and polycystic ovary syndrome (unpublished work).

To further enhance the ease of use, tutorial video material should be made available for the end users. Also, detailed production-ready workflows for many different scenarios, such as whole genome resequencing, RNA-seq analysis and GWAS workflow, should be provided with the software. Lastly, the flexibility of the software for workflow prototyping purposes could be enhanced by allowing users to define ranges for various parameters to be automatically tested.

3.2 Population structure within *P. sylvestris* main range is due to isolation-by-distance

In study I the analysis of genetic diversity Table 1 shows similar results to what have been found in other conifer species (Eckert, Bower, Jermstad, & Wegrzyn, 2013) and in earlier *P. sylvestris* studies (Pyhäjärvi et al., 2007; Kujala et al., 2017), with limited differences between populations. These values were calculated for fourfold degenerate sites which are, by definition, found within coding regions, but diversity values outside exons may be different (Andolfatto, 2005). Tajima's D estimates were negative indicating overrepresentation of rare variants. Based on the diversity estimates presented in this thesis, combined with mutation rate estimates and assuming a generation time of 20 years, the effective population size has been suggested to be 38,000 to 230,000 (Pyhäjärvi et al., 2020). This is a small number compared to the likely census size of several hundred billion, caused likely by the following violations of some of the assumptions: uneven fecundity of individual trees, a lower mutation rate than estimated from fossil record, nonequilibrium population history, and the effect of linked selection within genic areas. The exact contributions of each of the listed reasons are not known, but at least historical population size variation due to glaciation periods (Pyhäjärvi et al., 2007; Kujala & Savolainen, 2012) and background selection (De La Torre, Li, Van De Peer, & Ingvarsson, 2017) likely play a key role explaining the difference (Pyhäjärvi et al., 2020).

Very low pairwise F_{st} values and low percentage of variance explained by first principal components in PCA suggest minimal trace of population structure within the main, continuous, range of *P. sylvestris*. Close examination of PCA results indicates some differentiation between eastern and western sampling sites, with more a larger difference suggested by STRUCTURE. However, the conStruct analysis, a comparison of the fit of a STRUCTURE-like non-spatial model to a spatial model accounting for isolation-by-distance showed that the latter model explains the data better. Thus, the clustering indicated by STRUCTURE likely is spurious. Such a lack of population structure is an uncommon feature in tree species, as species with similar wide distributions, such as *Picea abies* and various *Populus* species, have been shown to exhibit more distinct population structure (Chen et al., 2019; De Carvalho et al., 2010; Geraldes et al., 2014; Keller, Levsen, Olson, & Tiffin, 2012). On the other hand, in North America *Pinus contorta* and interior spruce (*Picea glauca, Picea engelmannii* and their hybrid) have been shown exhibit limited population structure (Yeaman et al., 2016).

3.3 Population and landscape genomics methods reveal few loci contributing to local adaptation

The linear regression modelling approach and the bayescan analysis for uncovering loci contributing to local adaptation produced very few outliers. The PCAdapt approach identified a larger number of putative outlier loci, although the fact that the other methods identified notably fewer loci may suggest that part of these findings are spurious. On the other hand, applying different kinds of statistical tests addressing different aspects of the data often yields different results, even on the same dataset (Biswas & Akey, 2006). The often used bayescan is an F_{ST} outlier method, which has some underlying similarities to PCAdapt, but the latter shows good performance in identifying outliers when in a range expansion has occurred in the sampled area (Luu et al., 2017). Furthermore, the linear regression model possibly lacks power in detecting abrupt allele frequency changes, such as predicted by Barton’s (1999) models, as a linear relationship between dependent and explanatory variables is expected in an outlier in this test.

Further, these results are not surprising in the light of theoretical and empirical literature suggesting that the genetic basis of local adaptation may be highly polygenic, resulting in only small allele frequency changes at individual loci across the range (Boyle, Li, & Pritchard, 2017). Including larger numbers of samples and populations would most likely aid in detecting more of these loci. More powerful approaches exist, such as utilizing polygenic scores obtained from genome-wide association studies, but such approach an requires large sample sizes and study of the adaptive traits as well (Berg & Coop, 2014). Many of the loci responsible for the adaptation may lie outside the coding areas included in our exome capture target, or outside of coding areas altogether. A whole genome sequencing experiment would be required for pinpointing these loci, but such experiments are not yet economically feasible in conifers, at least in sufficiently large scale.

3.4 A putative large inversion may contribute to local adaptation

Theory and empirical findings suggests that local adaptation under gene flow favours a genomic architecture where underlying loci are concentrated in areas of reduced recombination (Yeaman & Whitlock, 2011; Samuk et al., 2017; Hämälä & Savolainen, 2019). Indeed, the outlier scans PCAdapt and Bayescan uncovered a haplotype structure spanning several hundred megabasepairs putatively under selection. Outlier scans are certainly prone to producing false positive signals (Hoban et al., 2016), but the main confounding factor of population structure appears to be very limited in *P. sylvestris*. Haplotype structures, namely inversions, are known to be under selection in many species (Lowry & Willis, 2010; Todesco et al., 2020; Wellenreuther & Bernatchez, 2018) but no such observations, to our knowledge, have been made in conifers. The lack of such observations in conifers and or in many other species may be partly due to the common tradition of using LD-pruning in filtering variants for selection scans but masking long range LD by default can hide these interesting features.

The cause for the haplotype structure cannot be confirmed from the dataset at hand, but inversion seems like the most plausible cause as no other explanation sufficiently describes how such a long stretch of the genome can be in full linkage with no recombination taking place between haplotypes. Many methods exist for detecting inversions (Kosugi et al., 2019), but are not applicable to the combination of fragmented reference genome and targeted sequence capture. The most plausible approach for testing for the presence of inversion would likely be to compare genetic maps generated from two different samples, one containing the inverted haplotype and the other without. Perhaps the emergence of long read sequencing will see wider use in resequencing studies in the future and enable better detection of structural variation. In the case of conifers it may also be, that structural variation is rare, as there appears to be high level of synteny between species (Pavy et al., 2012).

3.5 Comparison of *P. sylvestris* and *A. lyrata* studies

The goals in studies I and II overlap, as both aimed to investigate the population structure and genetic basis of adaptation, although study I was more aimed towards the latter goal and study II more towards the demographic and colonization history. Even though the study goals were similar, different approaches were used from study design, sequencing approach, bioinformatics workflow to the population genetics analysis.

Sampling in study I was designed to cover large part of the species distribution and to uncover the genetic basis of phenotypic clines by sampling along two latitudinal gradients. In study II the sampling design was aimed to facilitate the study of the post-glacial recolonization of Scandinavia. In study I, a targeted sequence capture was a suitable way to examine coding regions at a large scale, as the expense of whole genome sequencing is still rather limiting, although not entirely prohibitive (Wang, Bernhardsson, & Ingvarsson, 2020), when studying species with very large genomes. Pooled sequencing would not allow efficient detection and filtering of paralogous regions and other approaches such as RAD-seq would only grant access to seemingly random, often anonymous parts of the genome. In study II a whole genome sequencing could be applied due to the small size of the *A. lyrata* genome, although a large proportion of the genome had to be masked from subsequent analysis due to the presence of repetitive content.

Bioinformatics workflows differed between the two studies due to the differences in ploidy level of the materials, lacking genomic resources in *P. sylvestris* and different objectives in the two studies. Haploid DNA material extracted from *P. sylvestris* aided in detection of paralogous sequences, alleviated the need of high sequencing depth present with diploid data and eased the examination of haplotypes, but complicated the analysis of data due to many commonly used software not permitting the use of haploid data. Furthermore, the large reference genome made bioinformatics workflow run very long lasting many weeks and prevented the use of many commonly used tools, such as GATK. In study II the runtimes of the bioinformatics workflow were manageable, but the repetitive content still caused many issues visible in the allele frequency spectrum leading to more complicated filtering efforts, and possibly resulting in removing large proportion of valid data as well. The low read depth of many samples prohibited traditional variant calling, but the software ANGSD provided a workaround via permitting all analysis by the use of genotype likelihoods.

Population structure of both study species was relatively well-known from literature, with very little structure within the main distribution range of *P. sylvestris* and high differentiation between the patchy *A. lyrata* populations. In both cases PCA and pairwise F_{ST} were used to assess the population structure, with highly similar tools STRUCTURE (in study I) and ADMIXTURE (in study II). More careful analysis of structure was done in different ways. In the case of *P. sylvestris*, conStruct was used to investigate whether the finding of structuring within the main range is truly due to improved resolution provided by large number of samples and variants in the current study or could such findings be due to isolation-by-distance. In the case of *A. lyrata,* demographic modelling was applied to uncover details of colonization history. The effect on allele frequency spectrum shape resulting mainly from different population histories in these two samples from two different species is evident in Figure 2.

Selection scans were performed differently in the studies. In study I, the aim was to detect allele frequency clines and other unexpected allele frequency differences by using landscape and F_{ST} outlier methods. In study II, an F_{ST} outlier method, population branch statistic (PBS) (Yi et al., 2010) was used. As the targeted sequencing approach prevented the use of sliding window analysis in study I, a permutation approach was taken instead to demonstrate “peakness” in F_{ST}, d_{XY} and π in the putative inversion region.

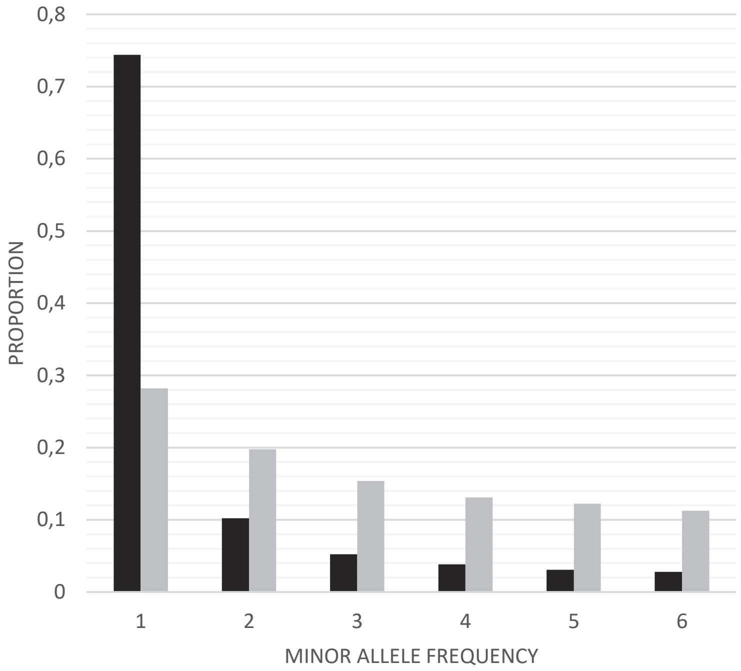


Fig. 2. Minor allele frequency spectra of all *P. sylvestris* samples (in black) and Plech population of *A. lyrata* (in grey) using data from studies I and II, respectively. *P. sylvestris* data downsampled to equal sample size to *A. lyrata* Plech population.

The two species studied in this thesis are outcrossing diploid plants, but they differ greatly in physiology, genome size and content, population sizes, histories and structure. These differences are then reflected in the choices made throughout the analysis workflow and interpretation of results. Nonetheless, interesting discoveries were made in both cases. As the generation time of *P. sylvestris* is roughly an order of magnitude longer compared to *A. lyrata,* the most recent glacial period is a lot more recent event for *P. sylvestris* from the perspective of selection. Therefore, some signature of adaptation to climate may still be relatively recent or in state of ongoing selective sweeps in *P. sylvestris*, but unfortunately due to the available sequencing approach in this species, the study of such signature is not possible. On the other hand, divergence-based methods for detecting selection in the genome are more straightforward in *P. sylvestris* due to the more subtle population structure.

4 Conclusions

There has been long-standing interest to better understand the genetic basis of local adaptation. Contemporary sequencing technologies have offered novel insights to this old question, but at the same time pose new difficulties to investigators in the form of computational challenges and new sources of bias in the data. In this thesis my aim was to study these themes by developing software to manage huge datasets, develop workflows to overcome technical issues and study genetic diversity to better understand genetic basis of local adaptation in *P. sylvestris.*

Our genome-wide dataset suggested some population structuring not seen before in the main range of *P. sylvestris,* but in-depth analysis revealed that most likely isolation-by-distance generated these results and no true structure exists. Theoretically, a species with a clear phenotypic signal of local adaptation in multiple contrasting environments with virtually non-existing population structure, such as *P. sylvestris,* is an ideal species for studying the genetic basis of local adaptation. Selection scans revealed some loci with interesting functions, but relatively few outlier loci were found, possibly due to the polygenic nature of adaptation and omitting non-coding regions from the study. Interestingly, a large putative inversion contributing to local adaptation was detected, a first such finding in conifers to our knowledge.

The whole genome sequencing data of *A. lyrata* was used for uncovering demographic history and post-glacial colonization patterns. The analysis relied on summary statistics largely affected by the shape of allele frequency spectrum, which contained significant bias due to technical errors. By applying careful bioinformatics workflow design and filtering approaches the bias was successfully reduced and further analysis was made possible.

I also developed a software "STAPLER" for managing and parallelizing bioinformatics workflows. The software allowed the prototyping of various alternative bioinformatics approaches for the studies of *P. sylvestris* and *A. lyrata*. When an optimal workflow was found, the parallelization features allowed a quick analysis of the whole data set. STAPLER has similar features to many other command-line workflow managers, but notably with built in flexibility and emphasis on ease-of-use. The results in this thesis emphasize the benefit of allocating time to bioinformatics workflow design. Even though an additional initial time investment is required, adopting tools to aid in the workflow design may be highly beneficial in speeding up the work and reducing possible bias in downstream population genetics analysis of the data.

The overall findings of genetic basis of local adaptation suggest that more emphasis is required for finding new approaches to detect polygenic adaptation. Furthermore, in scenarios with high geneflow the search of extended haplotype structures and structural variation can indeed be fruitful approaches.

List of references

Aho, M. (1994). Autumn frost hardening of one-year-old *Pinus sylvestris* (L.) seedlings: Effect of origin and parent trees. *Scandinavian Journal of Forest Research*, *9*(1–4), 17–24. https://doi.org/10.1080/02827589409382808

Alexander, D. H., Novembre, J., & Lange, K. (2009). *Fast model-based estimation of ancestry in unrelated individuals.* 1655–1664. https://doi.org/10.1101/gr.094052.109.vidual

Andoflatto, P., Depaulis, F., & Navarro, A. (2001). Inversion polymorphisms and nucleotide variability in Drosophila. *Genetics Research*, *77*(01), 1–8. https://doi.org/10.1017/S0016672301004955

Andolfatto, P. (2005). Adaptive evolution of non-coding DNA in Drosophila. *Nature*, *437*(7062), 1149–1152.

Barghi, N., Tobler, R., Nolte, V., Jakšić, A. M., Mallard, F., Otte, K. A., … Schlötterer, C. (2019). Genetic redundancy fuels polygenic adaptation in Drosophila. In *PLoS Biology* (Vol. 17). https://doi.org/10.1371/journal.pbio.3000128

Barton, N. H. (1999). Clines in polygenic traits. *Genetical Research*, *74*(3), 223–236. https://doi.org/10.1017/S001667239900422X

Barton, Nicholas H., & Keightley, P. D. (2002). Multifactorial Geneticsunderstanding Quantitative Genetic Variation. *Nature Reviews Genetics*, *3*(1), 11–21. https://doi.org/10.1038/nrg700

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., … Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, *456*(7218), 53.

Berg, J. J., & Coop, G. (2014). A Population Genetic Signal of Polygenic Adaptation. *PLoS Genetics*, *10*(8). https://doi.org/10.1371/journal.pgen.1004412

Beuker, E. (1994). Adaptation to climatic changes of the timing of bud burst in populations of Pinus sylvestris L. and Picea abies. *Tree Physiology*, *14*(7–9), 961–970.

Biswas, S., & Akey, J. M. (2006). Genomic insights into positive selection. *Trends in Genetics*, *22*(8), 437–446. https://doi.org/10.1016/j.tig.2006.06.005

Blanquart, F., Kaltz, O., Nuismer, S. L., & Gandon, S. (2013). A practical guide to measuring local adaptation. *Ecology Letters*, *16*(9), 1195–1205. https://doi.org/10.1111/ele.12150

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Booker, T. R., Yeaman, S., & Whitlock, M. C. (2019). Global adaptation confounds the search for local adaptation. *BioRxiv*, 742247. https://doi.org/10.1101/742247

Bourgey, M., Dali, R., Eveleigh, R., Chen, K. C., Letourneau, L., Fillon, J., … Bourque, G. (2019). GenPipes: An open-source framework for distributed and scalable genomic analyses. *GigaScience*, *8*(6), 1–11. https://doi.org/10.1093/gigascience/giz037

Boyle, E. A., Li, Y. I., & Pritchard, J. K. (2017). An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell*, *169*(7), 1177–1186. https://doi.org/10.1016/j.cell.2017.05.038

Bradburd, G. S., Coop, G. M., & Ralph, P. L. (2018). Inferring continuous and discrete population genetic structure across space. *Genetics*, *210*(1), 33–52.

Cheddadi, R., Vendramin, G. G., Litt, T., François, L., Kageyama, M., Lorentz, S., … Lunt, D. (2006). Imprints of glacial refugia in the modern genetic diversity of Pinus sylvestris. *Global Ecology and Biogeography*, *15*(3), 271–282. https://doi.org/10.1111/j.1466-822X.2006.00226.x

Chen, J., Li, L., Milesi, P., Jansson, G., Berlin, M., Karlsson, B., … Lascoux, M. (2019). Genomic data provide new insights on the demographic history and the extent of recent material transfers in Norway spruce. *Evolutionary Applications*, *12*(8), 1539–1551. https://doi.org/10.1111/eva.12801

Clarke, J., Wu, H.-C., Jayasinghe, L., Patel, A., Reid, S., & Bayley, H. (2009). Continuous base identification for single-molecule nanopore DNA sequencing. *Nature Nanotechnology*, *4*(4), 265.

Coop, G., Witonsky, D., Di Rienzo, A., & Pritchard, J. K. (2010). Using environmental correlations to identify loci underlying local adaptation. *Genetics*, *185*(4), 1411–1423. https://doi.org/10.1534/genetics.110.114819

Darlington, C. D., Mather, K., & others. (1949). The elements of genetics. *The Elements of Genetics*.

Darwin, C. (1859). *On the origin of species by means of natural selection, or, the preservation of favoured races in the struggle for life*. J. Murray.

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology: a comparison of two theoretical models. *Management Science*, *35*(8), 982–1003.

De Carvalho, D., Ingvarsson, P. K., Joseph, J., Suter, L., Sedivy, C., MACAYA-SANZ, D., … Lexer, C. (2010). Admixture facilitates adaptation from standing variation in the European aspen (Populus tremula L.), a widespread forest tree. *Molecular Ecology*, *19*(8), 1638–1650.

De La Torre, A. R., Li, Z., Van De Peer, Y., & Ingvarsson, P. K. (2017). Contrasting rates of molecular evolution and patterns of selection among gymnosperms and flowering plants. *Molecular Biology and Evolution*, *34*(6), 1363–1377. https://doi.org/10.1093/molbev/msx069

Dobzhansky, T. (1970). *Genetics of the evolutionary process* (Vol. 139). Columbia University Press.

Dohm, J. C., Lottaz, C., Borodina, T., & Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research*, *36*(16), e105.

Dvornyk, V., Sirviö, A., Mikkonen, M., & Savolainen, O. (2002). Low nucleotide diversity at the pal1 locus in the widely distributed Pinus sylvestris. *Molecular Biology and Evolution*, *19*(2), 179–188. https://doi.org/10.1093/oxfordjournals.molbev.a004070

Eckert, A. J., Bower, A. D., Jermstad, K. D., & Wegrzyn, J. L. (2013). *Multilocus analyses reveal little evidence for lineage-wide adaptive evolution within major clades of soft pines (Pinus subgenus Strobus)*. 5635–5650. https://doi.org/10.1111/mec.12514

Eiche, V. (1966). Cold Damage and Plant Mortality in Experimental Provenance Plantations with Scots Pine in Northern Sweden. *STUDIA FORESTALIA SUECICA*, *36*, 1–219.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., … Turner, S. (2009). Real-time DNA sequencing from single polymerase molecules. *Science*, *323*(5910), 133–138.

Excoffier, L., Dupanloup, I., Huerta-Sanchez, E., Sousa, V. C., & Foll, M. (2013). Robust Demographic Inference from Genomic and SNP Data. *PLoS Genetics*, *9*(10), e1003905. https://doi.org/10.1371/journal.pgen.1003905

Excoffier, L., & Ray, N. (2008). Surfing during population expansions promotes genetic revolutions and structuration. *Trends in Ecology and Evolution*, *23*(7), 347–351. https://doi.org/10.1016/j.tree.2008.04.004

Faria, R., Johannesson, K., Butlin, R. K., & Westram, A. M. (2019). Evolving Inversions. *Trends in Ecology and Evolution*, *34*(3), 239–248. https://doi.org/10.1016/j.tree.2018.12.005

Feldman, M. W., Otto, S. P., & Christiansen, F. B. (1997). Population Genetic Perspectives on the Evolution of Recombination. *Annual Review of Genetics*, *30*(1), 261–295. https://doi.org/10.1146/annurev.genet.30.1.261

Fisch, K. M., Meißner, T., Gioia, L., Ducom, J. C., Carland, T. M., Loguercio, S., & Su, A. I. (2015). Omics Pipe: A community-based framework for reproducible multi-omics data analysis. *Bioinformatics*, *31*(11), 1724–1728. https://doi.org/10.1093/bioinformatics/btv061

Fisher, R. (1918). The Correlation Between Relatives on the Supposition of Mendelian Inheritance. *Proc. Roy. Soc. Edinburgh*, *52*, 399–433.

Foll, M., & Gaggiotti, O. (2008). A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics*, *180*(2), 977–993. https://doi.org/10.1534/genetics.108.092221

Garrison, E., & Marth, G. (2012). *Haplotype-based variant detection from short-read sequencing*. arXiv:1207.3907. 1–20. [q-bio.GN]

Geraldes, A., Farzaneh, N., Grassa, C. J., McKown, A. D., Guy, R. D., Mansfield, S. D., … Cronk, Q. C. B. (2014). Landscape genomics of Populus trichocarpa: the role of hybridization, limited gene flow, and natural selection in shaping patterns of population structure. *Evolution*, *68*(11), 3260–3280.

Goecks, J., Nekrutenko, A., Taylor, J., & Galaxy Team, T. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, *11*(8), R86. https://doi.org/10.1186/gb-2010-11-8-r86

Gould, B. A., Chen, Y., & Lowry, D. B. (2018). Gene Regulatory Divergence Between Locally Adapted Ecotypes in Their Native Habitats. *Molecular Ecology*, (January), 1–15. https://doi.org/10.1111/mec.14852

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., & Bustamante, C. D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics*, *5*(10). https://doi.org/10.1371/journal.pgen.1000695

Haldane, J. B. S. (1930). A mathematical theory of natural and artificial selection. (Part VI, Isolation.). *Mathematical Proceedings of the Cambridge Philosophical Society*, *26*(02), 220. https://doi.org/10.1017/S0305004100015450

Hämälä, T., Mattila, T. M., & Savolainen, O. (2018). Local adaptation and ecological differentiation under selection, migration, and drift in Arabidopsis lyrata*. *Evolution*, *72*(7), 1373–1386. https://doi.org/10.1111/evo.13502

Hämälä, T., & Savolainen, O. (2019). Genomic Patterns of Local Adaptation under Gene Flow in Arabidopsis lyrata. *Molecular Biology and Evolution*, *36*(11), 2557–2571. https://doi.org/10.1093/molbev/msz149

Hoban, S., Kelley, J. L., Lotterhos, K. E., Antolin, M. F., Bradburd, G., Lowry, D. B., … Whitlock, M. C. (2016). Finding the Genomic Basis of Local Adaptation: Pitfalls, Practical Solutions, and Future Directions. *The American Naturalist*, *188*(4), 379–397. https://doi.org/10.1086/688018

Hu, T. T., Pattyn, P., Bakker, E. G., Cao, J., Cheng, J.-F., Clark, R. M., … Guo, Y.-L. (2011). The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nature Genetics*, *43*(5), 476–481. https://doi.org/10.1038/ng.807

Hudson, R. R., Slatkin, M., & Maddison, W. P. (1992). Estimation of levels of gene flow from DNA sequence data. *Genetics*, *132*(2).

Hurme, P., Repo, T., Savolainen, O., & Pääkkönen, T. (1997). Climatic adaptation of bud set and frost hardiness in Scots pine (*Pinus sylvestris*). *Canadian Journal of Forest Research*, *27*(5), 716–723. https://doi.org/10.1139/x97-052

Hurme, P., Sillanpää, M. J., Arjas, E., Repo, T., & Savolainen, O. (2000). Genetic basis of climatic adaptation in scots pine by Bayesian quantitative trait locus analysis. *Genetics*, *156*(3), 1309–1322. Retrieved from http://www.scopus.com/inward/record.url?eid=2-s2.0-0033766566&partnerID=tZOtx3y1

Jones, F. C., Grabherr, M. G., Chan, Y. F., Russell, P., Mauceli, E., Johnson, J., … Kingsley, D. M. (2012). The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, *484*(7392), 55–61. https://doi.org/10.1038/nature10944

Jung, H., Winefield, C., Bombarely, A., Prentis, P., & Waterhouse, P. (2019). Tools and strategies for long-read sequencing and de novo assembly of plant genomes. *Trends in Plant Science*.

Kapun, M., & Flatt, T. (2018). The adaptive significance of chromosomal inversion polymorphisms in Drosophila melanogaster. *Molecular Ecology*. https://doi.org/10.1111/mec.14871

Karhu, A., Hurme, P., Karjalainen, M., Karvonen, P., Kärkkäinen, K., Neale, D., & Savolainen, O. (1996). Do molecular markers reflect patterns of differentiation in adaptive traits of conifers? *Theoretical and Applied Genetics*, *93*(1–2), 215–221. https://doi.org/10.1007/s001220050268

Kawecki, T. J., & Ebert, D. (2004). Conceptual issues in local adaptation. *Ecology Letters*, *7*(12), 1225–1241. https://doi.org/10.1111/j.1461-0248.2004.00684.x

Keller, S. R., Levsen, N., Olson, M. S., & Tiffin, P. (2012). Local adaptation in the flowering-time gene network of balsam poplar, populus balsamifera L. *Molecular Biology and Evolution*, *29*(10), 3143–3152. https://doi.org/10.1093/molbev/mss121

Kircher, M., & Kelso, J. (2010). High-throughput DNA sequencing--concepts and limitations. *Bioessays*, *32*(6), 524–536.

Kirkpatrick, M., & Barton, N. (2006). *Chromosome Inversions, Local Adaptation and Speciation*. *434*(May), 419–434. https://doi.org/10.1534/genetics.105.047985

Korneliussen, T S, Moltke, I., Albrechtsen, a, & Nielsen, R. (2013). Calculation of Tajima's D and other neutrality test statistics from low depth next-generation sequencing data. *BMC Bioinformatics*, *14*, 289. https://doi.org/10.1186/1471-2105-14-289

Korneliussen, Thorfinn Sand, Albrechtsen, A., & Nielsen, R. (2014). ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics*, *15*(1), 356.

Köster, J., & Rahmann, S. (2012). Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics*, *28*(19), 2520–2522. https://doi.org/10.1093/bioinformatics/bts480

Kosugi, S., Momozawa, Y., Liu, X., Terao, C., Kubo, M., & Kamatani, Y. (2019). *Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing*. 8–11.

Kremer, A., & Le Corre, V. (2012). Decoupling of differentiation between traits and their underlying genes in response to divergent selection. *Heredity*, *108*(4), 375–385. https://doi.org/10.1038/hdy.2011.81

Kujala, S., Knürr, T., Kärkkäinen, K., Neale, D. B., Sillanpää, M. J., & Savolainen, O. (2017). Genetic heterogeneity underlying variation in a locally adaptive clinal trait in Pinus sylvestris revealed by a Bayesian multipopulation analysis. *Heredity*, *118*(5), 413–423. https://doi.org/10.1038/hdy.2016.115

Kujala, S., & Savolainen, O. (2012). Sequence variation patterns along a latitudinal cline in Scots pine (Pinus sylvestris): Signs of clinal adaptation? *Tree Genetics and Genomes*, *8*(6), 1451–1467. https://doi.org/10.1007/s11295-012-0532-5

Kumar, S., Tamura, K., & Nei, M. (1994). MEGA: molecular evolutionary genetics analysis software for microcomputers. *Bioinformatics*, *10*(2), 189–191.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Latta, R. G. (1998). Differentiation of Allelic Frequencies at Quantitative Trait Loci Affecting Locally Adaptive Traits Differentiation of Allelic Frequencies at Quantitative Trait Loci Affecting Locally Adaptive Traits. *The American Naturalist*, *151*(3), 283–292. https://doi.org/10.1086/286119

Latta, R. G. (2003). Gene flow, adaptive population divergence and comparative population structure across loci. *New Phytologist*, *161*(1), 51–58. https://doi.org/10.1046/j.1469-8137.2003.00920.x

Le Corre, V., & Kremer, A. (2003). Genetic variability at neutral markers, quantitative trait loci and trait in a subdivided population under selection. *Genetics*, *164*(3), 1205–1219. https://doi.org/10.1186/1471-2148-9-177

Le Corre, V., & Kremer, A. (2012). The genetic differentiation at quantitative trait loci under local adaptation. *Molecular Ecology*, *21*(7), 1548–1566. https://doi.org/10.1111/j.1365-294X.2012.05479.x

Leinonen, P. H., Remington, D. L., & Savolainen, O. (2011). Local adaptation, phenotypic differentiation, and hybrid fitness in diverged natural populations of arabidopsis lyrata. *Evolution*, *65*(1), 90–107. https://doi.org/10.1111/j.1558-5646.2010.01119.x

Leinonen, P. H., Sandring, S., Quilot, B., Clauss, M. J., Mitchell-Olds, T., Agren, J., & Savolainen, O. (2009). Local adaptation in European populations of Arabidopsis lyrata (Brassicaceae). *American Journal of Botany*, *96*(6), 1129–1137. https://doi.org/10.3732/ajb.0800080

Levins, R. (1968). *Evolution in changing environments; some theoretical explorations.* Princeton University Press. Retrieved from https://press.princeton.edu/titles/641.html

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, *26*(5), 589–595. https://doi.org/10.1093/bioinformatics/btp698

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, Z., Baniaga, A. E., Sessa, E. B., Scascitelli, M., Graham, S. W., Rieseberg, L. H., & Barker, M. S. (2015). *Early genome duplications in conifers and other seed plants*. (November), 1–8.

Lieberman-Aiden, E., Van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., … Dekker, J. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, *326*(5950), 289–293.

Liu, X., & Fu, Y.-X. (2015). Exploring population size changes using SNP frequency spectra. *Nature Genetics*, *47*(5), 555–559. https://doi.org/10.1038/ng.3254

Lowry, D. B., & Willis, J. H. (2010). A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation. *PLoS Biology*, *8*(9). https://doi.org/10.1371/journal.pbio.1000500

Lu, H., Giordano, F., & Ning, Z. (2016). Oxford Nanopore MinION sequencing and genome assembly. *Genomics, Proteomics & Bioinformatics*, *14*(5), 265–279.

Luu, K., Bazin, E., & Blum, M. G. B. (2017). pcadapt: an R package to perform genome scans for selection based on principal component analysis. *Molecular Ecology Resources*, *17*(1), 67–77. https://doi.org/10.1111/1755-0998.12592

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., … Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, *437*(7057), 376–380.

Mattila, T. M., Laenen, B., Horvath, R., Hämälä, T., Savolainen, O., & Slotte, T. (2019). Impact of demography on linked selection in two outcrossing Brassicaceae species. *Ecology and Evolution*, *9*(17), 9532–9545.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … DePristo M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303.

McKinney, G. J., Waples, R. K., Seeb, L. W., & Seeb, J. E. (2017). Paralogs are revealed by proportion of heterozygotes and deviations in read ratios in genotyping-by-sequencing data from natural populations. *Molecular Ecology Resources*, *17*(4), 656–669. https://doi.org/10.1111/1755-0998.12613

McVean, G. (2009). A genealogical interpretation of principal components analysis. *PLoS Genetics*, *5*(10). https://doi.org/10.1371/journal.pgen.1000686

Mikola, J. (1982). Bud-set phenology as an indicator of climatic adaptation of Scots pine in Finland [Pinus sylvestris]. *Population Genetics of Forest Trees*, *16*(2), 178–184.

Monat, C., Tranchant-Dubreuil, C., Kougbeadjo, A., Farcy, C., Ortega-Abboud, E., Amanzougarene, S., … Sabot, F. (2015). TOGGLE: toolbox for generic NGS analyses. *BMC Bioinformatics*, *16*(1), 374. https://doi.org/10.1186/s12859-015-0795-6

Muller, M.-H., Leppälä, J., & Savolainen, O. (2008). Genome-wide effects of postglacial colonization in Arabidopsis lyrata. *Heredity*, *100*(1), 47–58. https://doi.org/10.1038/sj.hdy.6801057

Naydenov, K., Senneville, S., Beaulieu, J., Tremblay, F., & Bousquet, J. (2007). Glacial vicariance in Eurasia: Mitochondrial DNA evidence from Scots pine for a complex heritage involving genetically distinct refugia at mid-northern latitudes and in Asia Minor. *BMC Evolutionary Biology*, *7*(1), 1–12. https://doi.org/10.1186/1471-2148-7-233

Neale, D. B., Wegrzyn, J. L., Stevens, K. A., Zimin, A. V, Puiu, D., Crepeau, M. W., … Langley, C. H. (2014a). Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. *Genome Biology*, *15*(3), R59. https://doi.org/10.1186/gb-2014-15-3-r59

Nei, M., Kojima, K.-I., & Schaffer, H. E. (1967). Frequency changes of new inversions in populations under mutation-selection equilibria. *Genetics*, *57*(4), 741.

Nei, M., & Li, W.-H. (1979). *Mathematical model for studying genetic variation in terms of restriction endonucleases*. *76*(10), 5269–5273.

Oinn, T., Addis, M., Ferris, J., Marvin, D., Senger, M., Greenwood, M., … Li, P. (2004). Taverna: A tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*, *20*(17), 3045–3054. https://doi.org/10.1093/bioinformatics/bth361

Ojeda, D. I., Mattila, T. M., Ruttink, T., Kujala, S. T., Kärkkäinen, K., Verta, J. P., & Pyhäjärvi, T. (2019). Utilization of Tissue Ploidy Level Variation in de Novo Transcriptome Assembly of Pinus sylvestris. *G3 (Bethesda, Md.)*, *9*(10), 3409–3421. https://doi.org/10.1534/g3.119.400357

Orr, H. A. (1998). The population genetics of adaptation: The distribution of factors fixed during adaptive evolution. *Evolution*, *52*(4), 935–949. Retrieved from http://www.scopus.com/inward/record.url?eid=2-s2.0-0031714655&partnerID=tZOtx3y1

Pavy, N., Lamothe, M., Pelgas, B., Gagnon, F., Birol, I., Bohlmann, J., … Bousquet, J. (2017). A high-resolution reference genetic map positioning 8.8 K genes for the conifer white spruce: structural genomics implications and correspondence with physical distance. *Plant Journal*, *90*(1), 189–203. https://doi.org/10.1111/tpj.13478

Pavy, N., Pelgas, B., Laroche, J., Rigault, P., Isabel, N., & Bousquet, J. (2012). A spruce gene map infers ancient plant genome reshuffling and subsequent slow evolution in the gymnosperm lineage leading to extant conifers. *Bmc Biology*, *10*(1), 84.

Pfeifer, S. P. (2017). From next-generation resequencing reads to a high-quality variant data set. *Heredity*, *118*(2), 111–124. https://doi.org/10.1038/hdy.2016.102

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945–959.

Puig, M., Villatoro, S., & Ca, M. (2015). *Human inversions and their functional consequences*. *14*(May), 369–379. https://doi.org/10.1093/bfgp/elv020

Pyhäjärvi, T, Garcia-Gil, M. R., Knurr, T., Mikkonen, M., Wachowiak, W., & Savolainen, O. (2007). Demographic history has influenced nucleotide diversity in European \emph{Pinus sylvestris} populations. *Genetics*, *177*(3), 1713–1724. https://doi.org/10.1534/genetics.107.077099

Pyhäjärvi, T., Aalto, E., & Savolainen, O. (2012). Time Scales of divergence and speciation among natural populations and subspecies of Arabidopsis lyrata (Brassicaceae). *American Journal of Botany*, *99*(8), 1314–1322. https://doi.org/10.3732/ajb.1100580

Pyhäjärvi, T., Hufford, M. B., Mezmouk, S., & Ross-Ibarra, J. (2013). Complex patterns of local adaptation in teosinte. *Genome Biology and Evolution*, *5*(9), 1594–1609. https://doi.org/10.1093/gbe/evt109

Pyhäjärvi, T., Kujala, S. T., & Savolainen, O. (2020). 275 years of forestry meets genomics in Pinus sylvestris. *Evolutionary Applications*, *13*(1), 11–30.

Rausher, M. D., & Delph, L. F. (2015). Commentary: When does understanding phenotypic evolution require identification of the underlying genes? *Evolution*, *69*(7), 1655–1664. https://doi.org/10.1111/evo.12687

Raymond, M. (1995). GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. *J. Hered.*, *86*, 248–249.

Rellstab, C., Gugerli, F., Eckert, A. J., Hancock, A. M., & Holderegger, R. (2015). A practical guide to environmental association analysis in landscape genomics. *Molecular Ecology*, *24*(17), 4348–4370. https://doi.org/10.1111/mec.13322

Ross-Ibarra, J., Wright, S. I., Foxe, J. P., Kawabe, A., DeRose-Wilson, L., Gos, G., … Gaut, B. S. (2008). Patterns of polymorphism and demographic history in natural populations of Arabidopsis lyrata. *PloS One*, *3*(6), e2411. https://doi.org/10.1371/journal.pone.0002411

Sadedin, S. P., Pope, B., & Oshlack, A. (2012). Bpipe: A tool for running and managing bioinformatics pipelines. *Bioinformatics*, *28*(11), 1525–1526. https://doi.org/10.1093/bioinformatics/bts167

Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., … Erlich, H. A. (1988). Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science*, *239*(4839), 487–491.

Saiki, R. K., Scharf, S., Faloona, F., Mullis, K. B., Horn, G. T., Erlich, H. A., & Arnheim, N. (1985). Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science*, *230*(4732), 1350–1354.

Samuk, K., Owens, G. L., Delmore, K. E., Miller, S. E., Rennison, D. J., & Schluter, D. (2017). Gene flow and selection interact to promote adaptive divergence in regions of low recombination. *Molecular Ecology*, *26*(17), 4378–4390. https://doi.org/10.1111/mec.14226

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, *74*(12), 5463–5467.

Savolainen, O., & Pyhäjärvi, T. (2007). Genomic diversity in forest trees. *Current Opinion in Plant Biology*, *10*(2), 162–167. https://doi.org/10.1016/j.pbi.2007.01.011

Savolainen, O., Pyhäjärvi, T., & Knürr, T. (2007). Gene Flow and Local Adaptation in Trees. *Annual Review of Ecology, Evolution, and Systematics*, *38*(1), 595–619. https://doi.org/10.1146/annurev.ecolsys.38.091206.095646

Shendure, J., Porreca, G. J., Reppas, N. B., Lin, X., McCutcheon, J. P., Rosenbaum, A. M., … Church, G. M. (2005). Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, *309*(5741), 1728–1732.

Skotte, L., Korneliussen, T. S., & Albrechtsen, A. (2013). Estimating individual admixture proportions from next generation sequencing data. *Genetics*, *195*(3), 693–702.

Slatkin, M. (1973). GENE FLOW AND SELECTION IN A CLINE. *Genetics*, *75*(4), 733–756.

Slatkin, M. (1978). Spatial patterns in the distributions of polygenic traits. *Journal of Theoretical Biology*, *70*, 213–280.

Smith, L. M., Sanders, J. Z., Kaiser, R. J., Hughes, P., Dodd, C., Connell, C. R., … Hood, L. E. (1986). Fluorescence detection in automated DNA sequence analysis. *Nature*, *321*(6071), 674–679.

Stevens, K. A., Wegrzyn, J. L., Zimin, A., Puiu, D., Crepeau, M., Cardeno, C., … Langley, C. H. (2016). Sequence of the sugar pine megagenome. *Genetics*, *204*(4), 1613–1626. https://doi.org/10.1534/genetics.116.193227

Swofford, D. L., & Selander, R. B. (1981). BIOSYS-1: a FORTRAN program for the comprehensive analysis of electrophoretic data in population genetics and systematics. *Journal of Heredity*, *72*(4), 281–283.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, *123*(3), 585–595. https://doi.org/PMC1203831

Todesco, M., Owens, G. L., Bercovich, N., Légaré, J.-S., Soudi, S., Burge, D. O., … Rieseberg L. H. (2020). Massive haplotypes underlie ecotypic differentiation in sunflowers. *Nature*, 1–6.

Wang, X., Bernhardsson, C., & Ingvarsson, P. K. (2020). Demography and natural selection have shaped genetic variation in the widely distributed conifer Norway Spruce (Picea abies). *Genome Biology and Evolution*, *12*(2), 3803–3817.

Wegrzyn, J. L., Liechty, J. D., Stevens, K. a, Wu, L.-S., Loopstra, C. a, Vasquez-Gross, H. a, … Neale, D. B. (2014). Unique features of the loblolly pine (Pinus taeda L.) megagenome revealed through sequence annotation. *Genetics*, *196*(3), 891–909. https://doi.org/10.1534/genetics.113.159996

Wellenreuther, M., & Bernatchez, L. (2018a). Eco-Evolutionary Genomics of Chromosomal Inversions. *Trends in Ecology & Evolution*, *33*(6), 427–440. https://doi.org/10.1016/j.tree.2018.04.002

Wellenreuther, M., & Bernatchez, L. (2018b). Eco-Evolutionary Genomics of Chromosomal Inversions. *Trends in Ecology and Evolution*, *33*(6), 427–440. https://doi.org/10.1016/j.tree.2018.04.002

Westbrook, J. W., Neves, L. G., Kirst, M., Peter, G. F., Chamala, S., Nelson, C. D., … Echt, C. S. (2015). A Consensus Genetic Map for Pinus taeda and Pinus elliottii and Extent of Linkage Disequilibrium in Two Genotype-Phenotype Discovery Populations of Pinus taeda . *G3&amp;#58; Genes|Genomes|Genetics*, *5*(8), 1685–1694. https://doi.org/10.1534/g3.115.019588

Wright, S. (1931). Evolution in Mendelian populations. *Genetics*, *16*(2), 97.

Yeaman, S., Hodgins, K. A., Lotterhos, K. E., Suren, H., Nadeau, S., Degner, J. C., … Aitken, S. N. (2016). Convergent local adaptation to climate in distantly related conifers. *Science*, *353*(6306), 1431–1433. https://doi.org/10.1126/science.aaf7812

Yeaman, S. (2015). Local Adaptation by Alleles of Small Effect. *The American Naturalist*, *186*(S1), S74–S89. https://doi.org/10.1086/682405

Yeaman, S, & Whitlock, M. C. (2011). The genetic architecture of adaptation under migration-selection balance. *Evolution; International Journal of Organic Evolution*, *65*(7), 1897–1911. https://doi.org/10.1111/j.1558-5646.2011.01269.x

Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z. X. P., Pool, J. E., … Wang, J. (2010). Sequencing of 50 human exomes reveals adaptation to high altitude. *Science (New York, N.Y.)*, *329*(5987), 75–78. https://doi.org/10.1126/science.1190371

Zimin, A., Stevens, K. a, Crepeau, M. W., Holtz-Morris, A., Koriabine, M., Marçais, G., … Langley, C. H. (2014). Sequencing and assembly of the 22-gb loblolly pine genome. *Genetics*, *196*(3), 875–890. https://doi.org/10.1534/genetics.113.159715

Zonneveld, B. J. M. (2012). Conifer genome sizes of 172 species, covering 64 of 67 genera, range from 8 to 72 picogram. *Nordic Journal of Botany*, *30*(4), 490–502. https://doi.org/10.1111/j.1756-1051.2012.01516.x

Original publications

I Tyrmi J.S., Vuosku J., Acosta J.J., Li Z., Sterk L., Cervera M.T., Savolainen O. & Pyhäjärvi T. (2020) Genomics of clinal local adaptation in *Pinus sylvestris* under continuous environmental and spatial genetic setting. G3: Genes | Genomes | Genetics.

II Mattila T.M., Tyrmi J., Pyhäjärvi T. & Savolainen O. (2017) Genome-Wide Analysis of Colonization History and Concomitant Selection in *Arabidopsis lyrata*. Molecular Biology and Evolution. 34(10):2665–2677.

III Tyrmi, J. (2020) STAPLER: a simple tool for creating, managing and parallelizing common high-throughput sequencing workflows. Manuscript.

Reprinted with permission from Genetics Society of America (I) and Oxford University Press (II).

Original publications are not included in the electronic version of the dissertation.

ACTA UNIVERSITATIS OULUENSIS

SERIES A SCIENTIAE RERUM NATURALIUM

732. Shevchuk, Nataliya (2019) Application of persuasive systems design for adopting green information systems and technologies

733. Tripathi, Nirnaya (2019) Initial minimum viable product development in software startups : a startup ecosystem perspective

734. Mohanani, Rahul Prem (2019) Requirements fixation: the effect of specification formality on design creativity

735. Salman, Iflaah (2019) The effects of confirmation bias and time pressure in software testing

736. Hosseini, Seyedrebvar (2019) Data selection for cross-project defect prediction

737. Karvonen, Juhani (2019) Demography and dynamics of a partial migrant close to the northern range margin

738. Rohunen, Anna (2019) Advancing information privacy concerns evaluation in personal data intensive services

739. Haghighatkhah, Alireza (2020) Test case prioritization using build history and test distances : an approach for improving automotive regression testing in continuous integration environments

740. Santos Parrilla, Adrian (2020) Analyzing families of experiments in software engineering

741. Kynkäänniemi, Sanna-Mari (2020) The relationship between the reindeer (*Rangifer tarandus tarandus*) and the ectoparasitic deer ked (*Lipoptena cervi*) : reindeer welfare aspects

742. Grundstrom, Casandra (2020) Health data as an enabler of digital transformation : a single holistic case study of connected insurance

743. Honka, Johanna (2020) Evolutionary and conservation genetics of European domestic and wild geese

744. Alasaarela, Mervi (2020) Tietojärjestelmän käytön vaikutus laatuun ja tuottavuuteen sairaalaorganisaatiossa palveluhenkilöstön kokemana

745. Korhonen, Tanja (2020) Tools and methods to support the key phases of serious game development in the health sector

746. Alahuhta, Kirsi (2020) Consequences of incomplete demographic information on ecological modelling of plant populations with hidden life-history stages

Book orders:
Virtual book store
http://verkkokauppa.juvenesprint.fi

UNIVERSITY OF OULU P.O. Box 8000 FI-90014 UNIVERSITY OF OULU FINLAND

ACTA UNIVERSITATIS OULUENSIS

UNIVERSITY OF OULU

ISBN 978-952-62-2734-4 (Paperback)
ISBN 978-952-62-2735-1 (PDF)
ISSN 0355-3191 (Print)
ISSN 1796-220X (Online)