

Report on Research Activities

A desire to understand the evolutionary process that creates organisms adapted to varying environmental conditions has been the driving force through my academic career. I have focused on plant population genomics tackling problems relating to local adaptation^{1,2}, effects of domestication and breeding^{3,4}, inference of population history⁵⁻⁷ and identifying the genetic footprint of natural selection in various ecological settings^{3,8,9}. The task is not easy. There are multiple processes simultaneously at work: random genetic drift, genetic variation along the genome, pleiotropy, spatial and temporal variation in environmental conditions and the ecological and genomic properties of plant species themselves. My research methodology relies strongly on understanding the theoretical basis of molecular variation: genetic drift, coalescent, mutation and other stochastic processes. Population genetic theory coupled with understanding of quantitative genetic variation is the foundation of all my research projects. My strength is the ability to combine modern large genomic datasets with the theoretical background and understanding the ecological context^{2,3,10}.

I have work on a number of different study species, each of which has unique challenges and questions. For example in teosinte and maize, genomic resources are highly developed, but population structure and hybridization may easily be confused with adaptive patterns. In contrast, most conifers are outcrossing species with large genomes and geographical distributions, and therefore are good species for understanding how adaptations to various environmental extreme conditions arise in biological settings that are not common in model species. Especially conifers thrill me, as they appear to have slow evolutionary rates and long generation time intervals, but have survived multiple glacial-interglacial cycles and demonstrated capability of rapid adaptation to changing conditions¹¹.

My research has clear translational implications for breeding, management and conservation. I have studied the impacts of breeding on maize genetic diversity in close collaboration with the USDA¹², estimated the adaptive capacity of Northern *P. sylvestris* populations under climate change¹³ and developed genomic resources in collaboration with forestry specialists of Natural Resources Institute Finland¹⁴. I have also studied genome-wide impacts of domestication, breeding and hybridization among wild and domestic lineages of maize and goose^{4,12}, as well as genetic composition of endangered Finnish wolf population¹⁵.

In the following sections, I highlight three examples of my research accomplishments in different stages of my career that are relevant for the position as a professor or assistant/associate professor, genomics-based plant science at the University of Helsinki.

Clinal variation in *P. sylvestris*

During my PhD at the University of Oulu, Finland (2003-2008) I was thrilled by the evolutionary mechanisms that underlie species' ability to cope with very harsh conditions. *Pinus sylvestris*, Scots pine, is known to have a large geographical and environmental distribution, and strong local adaptation with a considerable genetic component¹⁶. Ample experimental forestry data implied that especially in the northern part of distribution, it cannot

be latitudinally transferred more than 100 kilometers without serious fitness effects¹⁶. Adaptive inference required good understanding of population demographic history. Using Sanger resequencing data, I made one of the first population demographic inferences based on genetic diversity and coalescent simulations in plants⁵. We concluded that very ancient population bottlenecks that have taken place maybe even 1 million years ago can still have an effect on *P. sylvestris* genetic diversity, due to long generation interval and large effective population sizes.

More recently, the European Union 7th FP ProCoGen project (2011-2015) greatly improved conifer genomic resources and provided my research group with data and the chance to revisit the questions of the genetic basis of *P. sylvestris* clinal variation. Aided by the first pine reference genome (*Pinus taeda*), we executed exome capture to study the genome-wide level of genetic structure and adaptive variation. We sampled 120 individuals along two latitudinal transects, and searched for adaptive patterns by multiple methods: regression analysis, environmental association and classic F_{ST} outlier analysis. As a part of MSc Jaakko Tyrmi's graduate studies¹⁰, we found that 1) the genetic structure of *P. sylvestris* is very weak and follows the spatial isolation-by-distance-model rather than a discrete island-model, 2) the signal of adaptation as allele frequency differences is seldom observed and 3) there is a curious > 300 Mbp haplotype structure that shows unexpectedly high linkage disequilibrium and population specific patterns indicating a putative inversion polymorphism that has captured a proportion of adaptive variation, as predicted by theory. Polymorphic inversions reduce recombination significantly and thus such a large inversion will also have implications in Scots pine breeding.

In conclusion, the genetic basis of adaptive variation in *P. sylvestris* is still elusive, possibly concentrated on non-coding regulatory regions that have not yet been assessed or as some theory predicts, demonstrating itself via allele frequency covariance patterns. To tackle the latter option and the analytical challenge of joint analysis of thousands or even millions of variants, we are currently collaborating with statistician Mikko Sillanpää's group in University of Oulu to develop suitable graphical lasso methods for identifying these loci.

Local adaptation in teosinte

As a fellow of the Academy of Finland in UC Davis (2010-2012), I shifted my study focus on maize and teosinte, which had more ample genomic resources available. I examined the genetic basis of local adaptation in teosinte (*Zea mays* ssp. *parviglumis* and ssp. *mexicana*), two wild relatives of domesticated maize^{2,12,17,18}. Since the maize reference genome sequence was already available, these species allowed for more detailed understanding on the genomic distribution of adaptation. The dataset was large, consisting of 40 000 SNPs genotyped in 250 individuals from 21 populations. We combined the genetic data with environmental data from climate and soil databases and applied the classical divergence outlier and environmental correlation methods to identify loci and genomic regions underlying local adaptation.

As I had hoped, the genome-wide view on the adaptation revealed novel aspects of the adaptation². First, we identified multiple large inversions that seemed to have caught adaptive variation and differentiate among populations and subspecies. In another study that I coauthored, some of the same inversions seemed to have introgressed from wild teosinte to maize¹⁸. This finding was significant because it showed that crops can utilize novel variation found in locally adapted wild relative to adapt to novel climates.

Second, we observed that in contrast to humans and *A. thaliana*, loci displaying signs of local adaptation in teosinte were actually enriched outside coding regions of the genes². Since teosinte and maize genomes are large and complex, the finding indicated that the adaptation process may be fundamentally different in plants that have large complex genomes. This is significant because a large number of plant species have large genomes, and much of the research has concentrated on small-genome species whose results may not be widely generalizable. In large-genomes, the adaptation may potentially have a larger mutational space, extending to regulatory regions and transposable elements.

Comparative genomic analysis of forest tree local adaptation

While working as an Academy Research Fellow at the University of Oulu (2015-present), I have been a Task Leader (major responsibility of subproject) in two European Union H2020 funded forest genetics projects, GenTree (2016-2020) and B4EST (2018-2022). For B4EST, I also heavily participated in proposal preparation (total budget 6,000,000 €) and represented the University of Oulu in the related negotiations. Both of these projects' purpose is to use a scientific approach to fight the adverse effects of climate change on both natural and managed forest tree populations and the related economy. GenTree is more concentrated on providing knowledge, methods and tools for the management and sustainable use of forest genetic resources, whereas B4EST focuses on actions to increase forest survival and health under climate change and natural disturbances.

The purpose of the GenTree subproject that I lead is to understand how different tree species are genetically adapted to contrasting environments in a European-wide and local scale. Together with multiple European partners, we have collected samples from contrasting environments around Europe from seven angiosperm and gymnosperm tree species. We have used targeted DNA sequencing to focus the analysis on 1000 genes in key functional pathways. 500 individuals per species will allow us to measure genetic diversity, estimate demographic history and identify potential large-effect loci conferring local adaptation. We will use the data to study possible convergent patterns across species and will also evaluate how population size and current and past geographic distribution will affect the molecular basis of adaptation. The project is a multidisciplinary effort combining contributions from geneticists, forest tree breeders and ecological modelers.

Rapid response to new climate and economic conditions requires nimble breeding methods. Applying the modern genomic prediction tools has a huge potential for forest trees as many of them are still in the early stages of breeding and traditional breeding is slow. In B4EST, we are taking a step further into practical forest genetic applications. Importantly, we develop three public 50 K SNP chips to facilitate more efficient genotyping of the most economically important European forest tree species: one for each Norway Spruce and Scots pine, and one that will be shared among four forest tree species. My research group has been responsible for Scots pine SNP chip design and we have combined our Illumina RNA and DNA sequence polymorphism data with earlier Sanger sequencing data and PacBio derived data from a collaborator Sonja Kujala in Natural Resources Institute to identify most reproducible and informative SNPs. The work is still ongoing, but will allow fast, easy and affordable large scale genotyping of Scots pine, which is a prerequisite for the applying modern genomic prediction methods in forest trees. Later in the B4EST project, we intend to test various genomic prediction

training sets for multiple species and test model performances across environment and genetic material.

Additional current projects

My research group works on a number of related projects, focusing on forest tree adaptive variation and role of life-history traits in conifer evolution. Here I list a few current areas of active research.

- Investigating the effect of first cycle breeding on genetic diversity in Finnish *P. sylvestris* (GENOWOOD project, in collaboration with University of Helsinki and Natural Resources Institute Finland)
- Reciprocal common garden experiments to investigate early fitness components role in *P. sylvestris* local adaptation (GenTree project, in collaboration with INIA, Spain and NERC, UK)
- Investigating allele specific expression and imprinting in *P. sylvestris* embryos (Academy Research Fellow project)
- Investigating the role of haploid life stage in evolutionary dynamics of conifers (Academy Research Fellow project)
- Modelling and simulating the evolution of polyzygotic polyembryony as a mechanism for avoiding inbreeding depression (In collaboration with Yaniv Brandvain, University of Minnesota)
- Understanding the mechanisms of gene family evolution in host-parasite system in *Plantago* (In collaboration with Anna-Liisa Laine, University of Helsinki and University of Zurich)

References

1. Savolainen, O., Pyhäjärvi, T. & Knürr, T. Gene flow and local adaptation in trees. *Annu. Rev. Ecol. Evol. Syst.* **38**, 595–619 (2007).
2. Pyhäjärvi, T., Hufford, M. B., Mezouk, S. & Ross-Ibarra, J. Complex patterns of local adaptation in teosinte. *Genome Biol. Evol.* **5**, 1594–1609 (2013).
3. Hufford, M. B. *et al.* Comparative population genomics of maize domestication and improvement. *Nat. Genet.* **44**, 808–811 (2012).
4. Heikkinen, M. E. *et al.* Relationship between wild greylag and European domestic geese based on mitochondrial DNA. *Anim. Genet.* **46**, (2015).
5. Pyhäjärvi, T. *et al.* Demographic history has influenced nucleotide diversity in European *Pinus sylvestris* populations. *Genetics* **177**, 1713–1724 (2007).
6. Pyhäjärvi, T., Salmela, M. J. & Savolainen, O. Colonization routes of *Pinus sylvestris* inferred from distribution of mitochondrial DNA variation. *Tree Genet. Genomes* **4**, 247–254 (2008).
7. Mattila, T. M., Tyrmi, J. S., Pyhäjärvi, T. & Savolainen, O. Genome-wide analysis of colonization history and concomitant selection in *Arabidopsis lyrata*. *Mol. Biol. Evol.* **34**, 2665–2677 (2017).
8. Toivainen, T., Pyhäjärvi, T., Niittyvuopio, A. & Savolainen, O. A recent local sweep at the PHYA locus in the Northern European Spiterstulen population of *Arabidopsis lyrata*. *Mol. Ecol.* **23**, 1040–52 (2014).
9. Fang, Z. *et al.* Megabase-scale inversion polymorphism in the wild ancestor of maize.

- Genetics* **191**, 883–894 (2012).
10. Tyrmi, J. *et al.* Genomics of clinal local adaptation in *Pinus sylvestris* under continuous environmental and spatial genetic setting. *bioRxiv* 647412 (2019).
 11. Savolainen, O., Pyhäjärvi, T. & Knürr, T. *Gene flow and local adaptation in trees. Annual Review of Ecology, Evolution, and Systematics* **38**, (2007).
 12. Hufford, M. B. *et al.* Comparative population genomics of maize domestication and improvement. *Nat. Genet.* **44**, (2012).
 13. Savolainen, O. *et al.* Adaptive potential of northernmost tree populations to climate change, with emphasis on Scots pine (*Pinus sylvestris* L.). *J. Hered.* **102**, 526–536 (2011).
 14. Ojeda, D. I. *et al.* Utilization of Tissue Ploidy Level Variation in de Novo Transcriptome Assembly of *Pinus sylvestris*. *G3 Genes | Genomes | Genetics* **9**, 3409–3421 (2019).
 15. Niskanen, a. K. *et al.* Balancing selection and heterozygote advantage in major histocompatibility complex loci of the bottlenecked Finnish wolf population. *Mol. Ecol.* **23**, 875–889 (2014).
 16. Pyhäjärvi, T., Kujala, S. T. & Savolainen, O. 275 years of forestry meets genomics in *Pinus sylvestris*. *Evol. Appl.* **In Press**, (2019).
 17. Hufford, M. B., Bilinski, P., Pyhäjärvi, T. & Ross-Ibarra, J. Teosinte as a model system for population and ecological genomics. *Trends Genet.* **28**, 606–615 (2012).
 18. Hufford, M. B. *et al.* The Genomic Signature of Crop-Wild Introgression in Maize. *PLoS Genet.* **9**, (2013).