Project plan

Population structure and patterns of adaptation in birch

December 10, 2021

Contents

1 Participants

Project supervisor: Martin Lascoux
Project co-supervisor: Jennifer James
Subject reader: Mattias Jakobsson
Examiner: Pascal Milesi
Student: Janek Sendrowski

2 Background

Observed patterns of genetic variability and structure in present day populations are shaped by a complex interplay of many factors, including individual dispersal ability, environmental conditions, and past demographic events. Climatic events such as glaciations are particularly noteworthy: they had a broad impact on many species, leading to reductions in population size and range contractions [5]. Many species then gradually recolonized areas as global temperatures increased [9]. The route that individuals took in recolonizing, the distance from the refuge area, and the uneven inheritance of ancestral variation can all lead to patterns in population-level genetic data [5].

However, some of the genetic differences we observe between populations could also be the result of adaptation. For example, after a glaciation, as temperatures increase and species can recolonize previously inhospitable areas, they may encounter different climatic zones. This may result in species adapting to local environmental conditions [6]. Maintaining such adaptation can also lead to persistent population structure.

Gaining a better understanding of the complex factors that determine patterns of polymorphism within species is the broad aim of this project. This research will focus on tree species in Europe, particularly birch. Trees are a keystone species with limited dispersal ability; as sessile organisms, they are not able to migrate in the face of unfavorable climatic conditions, and thus might experience strong selective pressure to adapt to local environmental conditions. Birch trees have been the focus of a recent sequencing effort, with individual trees sampled in Sweden over a latitudinal gradient that includes a transition between different climatic zones. Whole genomes have also been sequenced from individuals sampled over the broad range of birch trees in Europe.

3 Project Importance

The project is of general importance to the understanding of the population structure and demographic history of birch and will contribute understanding (in the form of written reports), and an analysis pipeline that will be made publicly available on GitHub, both of which will be useful to researchers, particularly to collaborators and others in the field.

In the longer term, the project is important from a social perspective: studying past patterns of migration and adaptation of large, slowly growing, slowly dispersing species could help our understanding of what is likely to happen to such species as climate conditions change in future.

4 Objectives

The goals of the project can be broken down into the following steps:

Objective 1: Examine geographic differentiation, migration and isolation by distance.

- Determine whether there exist geographic differences.
- Estimate migration.
- Compare results with established results from literature.
- Potentially use genome-wide genealogy methods for estimating isolation by distance.

Objective 2: Estimate the distribution of fitness effects (DFE).

- Estimate the DFE for all *B. pendula* samples as a whole.
- Compare the DFE between populations in different climate zones.
- Include the estimation of the distribution of beneficial mutations.

Objective 3: Fit a model of the demographic history of the populations to the data.

- Review literature to identify preliminary demographic models to test.
- Consider the use of new methods - including genome-wide genealogy methods for estimating population dispersal rates.
- Check for accordance between inferred demography and results.

5 Practical

5.1 Procedure

The brunt of the work will be concerned with building the pipeline to perform the various analyses. Technical problems are thus likely to arise which can slow down the progress significantly. An efficient, highly parallelized and fully automatic pipeline is to be implemented using *snakemake* for workflow management. The use of this software ensures reproducibility as well as extensibility. The sample size can, for example, easily be increased later on to make the analyses more robust or diverse. Despite using UPPMAX for all computations, the pipeline will be platform-independent, solely relying on Conda for its dependencies.

The initial raw data consist of resequenced Illumina short reads primarily from *Betula pendula* and *B. pubescens* samples across Sweden and Norway. We will mostly focus on *B. pendula* in this work whose reference genome we use. The data need to be trimmed, checked for quality and mapped before the SNPs can be called. These analyses will be performed with Trimmomatic, FastQC, BWA and GATK, respectively. The reference genome as well as the alder outgroup species were obtained from a repository associated with a Finnish genome sequencing project [6]. Different filtering criteria are to be tested for the called variants to see which ones are most appropriate.

Initially, a PCA plot and a PCA-like tool called UMAP [4] will be used to provide information on how to subdivide the data in further analyses. A more sophisticated approach for clustering might also be appropriate. We calculate some more basic summary statistics like F_{ST} and π before proceeding with the estimation of migration surfaces with FEEMS [3]. We then compare the distribution of fitness effects (DFE) among different populations using polyDFE. For this we first need to distinguish synonymous from non-synonymous sites. If we additionally like to infer the distribution of beneficial mutations, we need to obtain an unfolded site frequency spectrum (SFS). We can do this by identifying the ancestral variant of our SNPs using an outgroup. After that, δaδi [2], which is also based on the SFS, will be used to test for the likelihood of different demographic models. DILS could subsequently be used to infer demographic scenarios as well, if time permits [1]. The formatting of input data seems to be rather demanding for this tool, however. The use of genome-wide genealogy methods might not be practical as the raw data is unphased. One idea is trying to phase it in retrospect, using statistical methods, but this may not provide results reliable enough.

5.2 Expectations

We generally do not expect a high level of population differentiation between different populations which is partly due to the large dispersal distances of the pollen. We do, however, expect differentiation along a latitudinal cline. The sample set we will work with also includes data from *B. pubescens* and other species which may have to be excluded for some analyses. We furthermore don't expect the DFE to differ between populations in different climate zones–despite their genetic differentiation. We hope to corroborate some results obtained from previous papers on birches [10, 5].

5.3 Meetings

Meetings will be held weekly. The student will also work on site so that possible issues can be addressed immediately.

5.4 Problems

If problems arise, the student will have to adjust the scope of the thesis, potentially reducing the number of analyses.

5.5 Schedule

September 1	Broad literature review
September 15	Familiarization with data set
September 22	Trimming, mapping and variant calling
October 1	PCA, UMAP and FEEMS [4, 3]
October 12	polyDFE [8]
October 24	δaδi [2]
Early November	(DILS, Relate [1, 7])
Mid November	Begin preparation of midterm presentation
Late November	Midterm presentation
Early December	Begin writing report
One week before final presentation	Submission of written report
Late January / early February	Final presentation

References

[1] Christelle Fraïsse, Iva Popovic, Clément Mazoyer, Bruno Spataro, Stéphane Delmotte, Jonathan Romiguier, Etienne Loire, Alexis Simon, Nicolas Galtier, Laurent Duret, Nicolas Bierne, Xavier Vekemans, and Camille Roux. Dils: Demographic inferences with linked selection by using abc. *Molecular Ecology Resources*, 01 2021.

[2] Ryan Gutenkunst, Ryan Hernandez, Scott Williamson, and Carlos Bustamante. Gutenkunst rn, hernandez rd, williamson sh, bustamante cd. inferring the joint demographic history of multiple populations from multidimensional snp data. plos genet 5: e1000695. *PLoS genetics*, 5:e1000695, 10 2009.

[3] Joseph Marcus, Wooseok Ha, Rina Barber, and John Novembre. Fast and flexible estimation of effective migration surfaces. *eLife*, 10, 07 2021.

[4] Leland McInnes and John Healy. Umap: Uniform manifold approximation and projection for dimension reduction. 02 2018.

[5] Anna Palme, Qiao Su, Anja Rautenberg, F Manni, and Martin Lascoux. Postglacial recolonization and cpdna variation of silver birch, betula pendula. *Molecular ecology*, 12:201–12, 02 2003.

[6] Jarkko Salojarvi, Olli-Pekka Smolander, Kaisa Nieminen, Sitaram Rajaraman, Omid Safronov, Pezhman Safdari, Airi Lamminmäki, Juha Immanen, Tianying Lan, Jaakko Tanskanen, Pasi Rastas, Ali Amiryousefi, Balamuralikrishna Jayaprakash, Juhana Kammonen, Risto Hagqvist, Gugan Eswaran, Viivi Hassinen, Juan Alonso-Serra, Fred Asiegbu, and Jaakko Kangasjärvi. Genome sequencing and population genomic analyses provide insights into the adaptive landscape of silver birch. *Nature Genetics*, 49, 05 2017.

[7] Leo Speidel, Marie Forest, Sinan Shi, and Simon Myers. A method for genome-wide genealogy estimation for thousands of samples. *Nature Genetics*, 51:1321–1329, 09 2019.

[8] Paula Tataru and Thomas Bataillon. *polyDFE: Inferring the Distribution of Fitness Effects and Properties of Beneficial Mutations from Polymorphism Data*, volume 2090, pages 125–146. 01 2020.

[9] Camille Truong, Anna Palme, and François Felber. Recent invasion of the mountain birch betula pubescens ssp. tortuosa above the treeline due to climate change: Genetic and ecological study in northern sweden. *Journal of evolutionary biology*, 20:369–80, 02 2007.

[10] Yoshiaki Tsuda, Vladimir Semerikov, Federico Sebastiani, Vendramin Giovanni Giuseppe, and Martin Lascoux. Multispecies genetic structure and hybridization in the betula genus across eurasia. *Molecular Ecology*, 26, 10 2016.