# Pangenome of silver birch: unlocking the hidden potential of structural variation for tree adaptation and breeding

## 1. Description of the proposed activities – MAX 5 PAGES

*Summary* While profiling genetic variation in the form of single nucleotide polymorphisms (SNPs) continues to offer excellent insight into population genetic variation, similar variation in genome structure (structural variation, SV), which may be equally important, has gone understudied. Advances in, and decreasing cost of, genome assembly offers a new approach to study SV by combining multiple genomes into a pangenome. Silver birch (*Betula pendula*), as an economically and ecologically important species, and with its small relatively small genome is an excellent candidate for a pangenome. Inspired by SV identified in the genome of a single birch individual, we propose work to identify and catalogue more of the SV present across the species range by assembling haplotype-resolved genomes from a larger number of individuals from across Europe. The resulting pangenome will shed light on the potential importance of large SV in local adaptation and tree breeding, as well as offering ways to use conventional short read data more effectively. As a pan-European project, we will bring together material and expertise from across the continent, creating a resource that is sure to be of immense value to the EVOLTREE community.

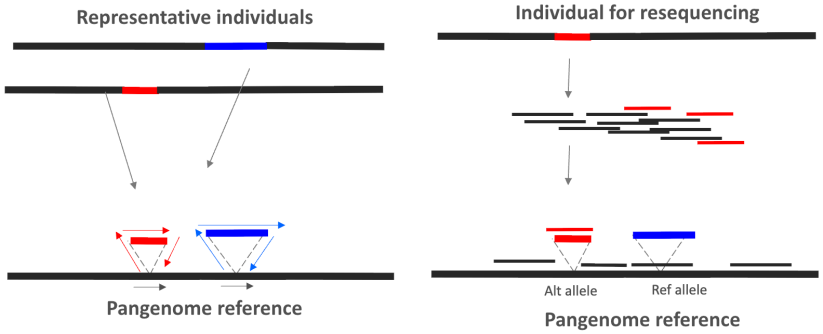


**Fig. 1** Schematic of pangenome graph and mapping of short resequencing reads. In this example of a reference pangenome graph (left), two genome-assembled individuals contain SVs relative to the reference individual – red and blue, respectively – which create 'bubbles' in the graph. Short reads from a resequenced individual can then be mapped to this graph (right). In this example, the resequenced individual carries the alternate allele for the red SV but the reference allele for the blue SV. Schematic by Daniel Wood.

*What is a pangenome?* While a reference genome represents a single set of chromosomes each with a single linear sequence, a pangenome represents a collection of genome assemblies from multiple distinct individuals, ideally capturing representative SV within the species. This reveals sets of genes, regulatory elements, or transposable elements that are either shared across the species ('core' genome) and those that are found only in some individuals ('accessory' genome). The graph-based pangenome data structure is comprised of a series of nodes and edges which capture the core genome sequence and the SVs that exist between individuals (**Fig. 1**, left). Recently developed tools allow mapping of short reads to a pangenome graph, allowing genotyping of structural variants which would otherwise have eluded detection if mapped to a linear reference genome (**Fig. 1**, right). Such structural variants not only provide markers of

genetic diversity complementary to SNPs but can also have powerful regulatory and phenotypic influences (e.g. coat pigmentation in corvids; Weissensteiner et al 2020).

A pangenome of 69 Arabidopsis accessions was found to comprise a total of around 33k genes compared to around 28k in the TAIR10 reference (Lian et al 2024). This fascinating diversity in gene repertoires poses both opportunities and challenges. On the one hand, 'accessory genes' (i.e. those found only in some varieties) offer another form of genomic diversity that may contribute to local adaptation and resilience, and which may be incorporated into genomic breeding efforts. For example, the pangenome of ash (*Fraxinus excelsior*) revealed over 3000 putative accessory genes, 133 of which have functions relevant to disease susceptibility (Wood et al 2025). Meanwhile, wild barley varieties were found to contain a larger gene repertoire than domesticated barley, illustrating substantial, and potentially useful variation missing in standard cultivars (Feng et al 2025). On the other hand, they can result in spurious patterns when aligning short reads to a single reference genome that does not contain the extensive gene repertoire. In *Arabidopsis*, copy number variants present in the sample but not in the reference result in multiple different gene copies mapping to the same gene in the reference, in turn resulting in spurious readouts of, e.g. DNAme from BSseq data (Jaegel et al 2023). Such spurious mappings of short reads can be avoided using a pangenome reference encapsulating more of the genomic structural variation (Igolkina et al 2025).

*Towards a birch pangenome* Birch (*Betula*) is a pioneer and keystone tree species of great ecological, economic, and cultural importance in Europe. With a wide European distribution, and as a fast-growing species that can be induced to flower within one year, and can be clonally propagated in vitro or by grafting, it is an excellent model for studying reproduction, development, and genetics in trees. Moreover, the relatively small genome size (440Mb) renders it attractive for large scale genomic analyses.

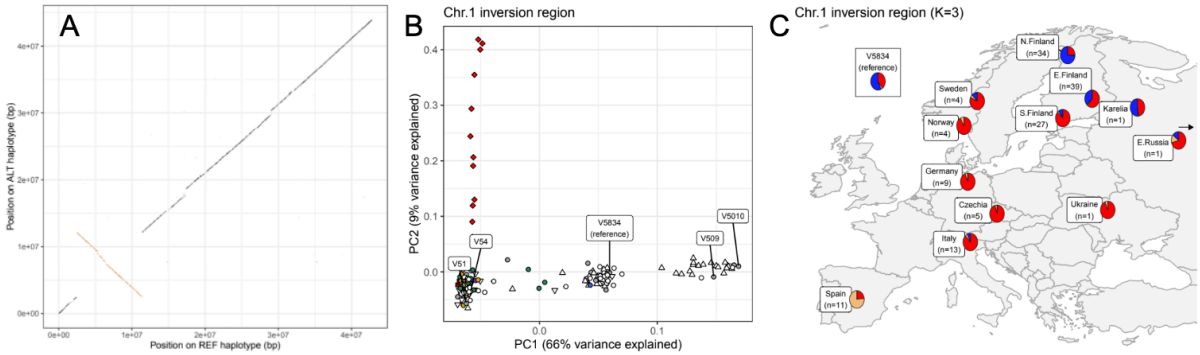


**Fig. 2**. Discovery of a large inversion and its segregation in European *B. pendula* populations. **(A)** After assembling both haplotypes of chromosome 1 from the reference individual V5834, alignment of the alternate haplotype (Y-axis) against the reference haplotype (X-axis) reveals a large inversion identified as heterozygous in the reference genome individual. **(B)** Short read sequencing of >170 individuals from across Europe followed by SNP calling and PCA of variants within the inversion region shows three distinct clusters corresponding with two homozygous genotypes and heterozygotes. **(C)** Admixture plot shows the haplotype composition of trees or populations of trees at different locations, with the alternate inversion haplotype (blue) being overrepresented at higher latitude. Figures from Ord et al (*in prep*).

The first *Betula pendula* was published in 2017 by UH researchers led by Jarkko Salojärvi (Salojärvi et al, 2017). The first genome provided extensive representation of the gene repertoire but it was not possible to resolve variation in genome structure. More recently, we assembled a new version of the *B. pendula* genome with PacBio HiFi sequencing. With this it was possible to resolve both haplotypes of the heterozygous chromosome set. Aligning these haplotype sequences against each other revealed that a long stretch of chromosome 1 was inverted between the two haplotypes (**Fig. 2**). Moreover, SNPs within the region, detected from short read sequencing data from dozens of individuals, revealed three genotypes present in the broader population (corresponding with two homozygous genotypes and heterozygotes). The distribution of genotypes even showed a geographic pattern, with the alternate haplotype being most prevalent in the north and completely absent in the southern latitudes, suggesting possible associations with cold tolerance and / or phenology. Individuals harbouring different inversion genotypes also show differences in vertical vs horizontal growth (data not shown) This analysis is in its final stages (Ord et al, in prep).

The discovery of the large inversion in the reference individual, V5834, was thanks to the capacity of read sequencing and genome assembly methods to resolve both haplotypes of a heterozygous individual. Other, complementary approaches however have revealed that SV is extensive in birch populations. Namely, genome assembly from short read data, which results in fragmented assemblies, nevertheless indicates a large degree of gene presence-absence variation (Rajaraman et al, in prep). Furthermore, PacBio data from additional individuals generated by the Nieminen Lab at UH (though with fragment lengths mostly not suitable for genome assembly) identified another, less common large inversion on chromosome 8 that also segregated in the population (data not shown). Finally, a smaller putative inversion on chromosome 6 was recently shown to be segregating in UK birch populations (Carleial et al 2025) and was associated with height and diameter. SV is therefore prevalent across the species range and may well have phenotypic consequences relevant to breeding and local adaptation, and there is therefore a need for more extensive characterisation of natural variation in birch genome structure. For this, high quality genome assemblies generated from long read sequence data of multiple individuals are needed. Thus, the direction of the existing research naturally points us towards a pangenome.

*Objectives* Our objective is to **generate high quality genome assemblies of several *B. pendula* individuals from across the species' European range**, capturing common large structural variants and the repertoires of core and accessory genes into a **pangenome**. As well as describing the extent of this structural variation, we will also **test the usefulness of the birch pangenome as a reference** with which to genotype SVs from short read data for applications such as GWAS, and to potentially improve the accuracy of short read-based population genomic and epigenomic analyses. At the University of Helsinki greenhouse, we have a collection of living birch specimens from across Finland and the rest of Europe (mostly from grafts) ranging from Northern Finland to Greece and including also specimens from Northern Italy, Sweden, Czechia, France, and Ukraine, and Poland (**Fig. 3**). These were mostly collected as part of a project previously set up by Jarkko Salojärvi to profile genetic variation across the European range (analysed as part of the new birch genome paper; Ord et al, in prep). Also growing on-site are specimens of idiosyncratic varieties such as the naturally occurring curly

birch which, although rarer, nevertheless occurs naturally in Finland. To these we will add specimens from Romania and Moldova with the assistance of **Dragos Postolache** (National Institute for Research and Development in Forestry, Romania), and specimens from two GenTree sites (Milesi et al, 2024) – Lithuania and Northern Spain, respectively. Although rare, the species also occurs in Northern Portugal, from where a specimen will be acquired with the assistance of **Filipe Costa e Silva** (Universidade de Lisboa Instituto Superior de Agronomia). To a total of 16 individuals which we plan to sequence, we will add already-generated high-quality assemblies from at least two others: our new reference genome individual V5834 of Finnish origin, and a sequenced individual of Scottish origin assembled as part of the Darwin Tree of Life Project. After assembling the genomes of all individual specimens, they will be annotated for protein coding genes and combined into a pangenome graph, enabling downstream analyses to address both biological and more technical questions.

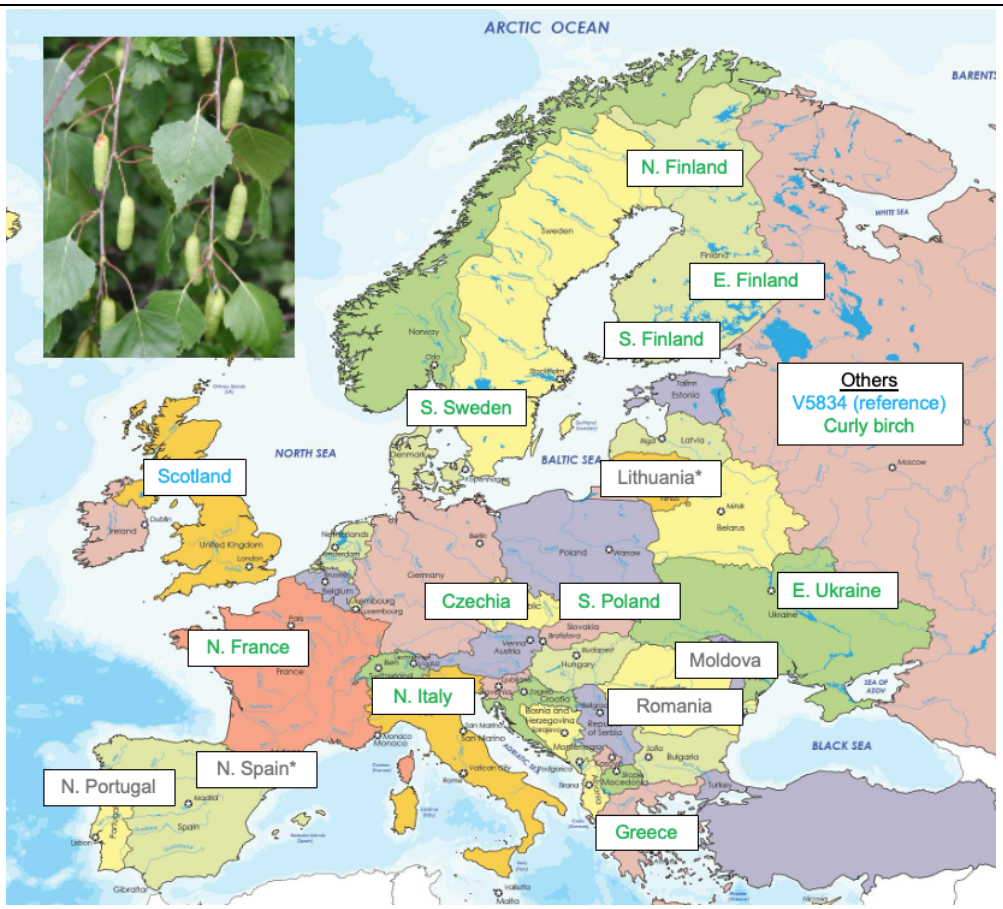


**Fig. 3. Sampling scheme**. Locations in green text have representative living specimens at the Viikki campus, UH. Locations in grey text have not been sampled but natural stands will be sampled with the help of collaborators (* denotes an established GenTree site, see Milesi et al 2024). Blue text denotes a location or accession for which a chromosome-level genome assembly is already available (either published or unpublished) to be incorporated into the pangenome.

*Downstream analyses* Initially, we will characterise our pangenome by asking how many unique genes in total are identified across all individual vs the number of genes common to all individuals (i.e. accessory vs core genes). We will then ask: **does accessory gene content of range edge specimens differ from range core specimens?** Such divergence

in gene presence absence variation may result from local adaptation, reduced gene flow (which may result in higher prevalence of deleterious gene deletions), or introgression from other species (e.g. *B. pubescens* or *B. nana* in the north). Concurrently, we will test the pangenome graph as a reference against which to genotype larger numbers of individuals sequenced with short reads (see **Fig. 1**, right). On a basic level, we can ask if a higher proportion of reads can be aligned to a pangenome reference compared to a single linear reference, thus allowing more biological information to be obtained from short read sequencing (or if the mapping quality can be improved, see e.g. Wood et al 2025). This will be tested using both whole genome resequencing reads from different populations (some of which are already published, see Salojärvi et al 2017), as well as a smaller amount of whole genome bisulfite sequencing reads. We will also however test if structural variants can be genotyped from short reads aligned to the pangenome, and whether such SV genotyping reveals biologically relevant information that might otherwise be missed (see e.g. Zhou et al 2022). Namely, we will perform a GWAS analysis of around 410 birch individuals from Eastern Finland (mapping population project established by Tanja Pyhäjärvi) to link genomic variants with measured phenotypic traits in the progeny. Key questions here are (1) whether more significant trait-associated SNPs are identified via the pangenome approach vs the linear genome approach and (2) the extent to which SVs vs SNPs are associated with phenotypic traits.

*Expected outcome* The pangenome will be an extremely useful asset for the community, unlocking the potential diversity of large polymorphisms. Promising potential uses include, for example: (1) testing the association between SV and desirable traits (such as pathogen resistance) in GWAS-like analyses, (2) checking whether genes of interest are duplicated, deleted, or fall within large inversions, (3) analysing the role of SV in local adaptation, and (4) increasing the accuracy of short read-based experiments (WGS, RNAseq, and BSseq) by mapping short reads to the pangenome graph instead of a linear reference genome. Eighteen individuals is a more comprehensive sampling than other recent pangenomes (e.g. cucmber, tea, millet; see Wood et al 2025, Table S6), however it may not capture all segregating SVs or accessory genes in *B. pendula* (indeed this was not found in the *Fraxinus excelsior* pangenome). Nevertheless, this resource will allow characterisation of thousands of common SVs segregating across Europe and will lay a foundation to which future genome assemblies can be added. As climate change intensifies the biotic and abiotic threats faced by natural populations, a high-quality catalogue of genomic SV of putative adaptive value has a chance to bolster efforts to counter emerging threats to populations (e.g. pests and pathogens).

*References*


Carleial, R., et al, 2025. *bioRxiv*, pp.2025-07.
Feng, J.W., et al, 2025. *Nature*, *645*(8080), pp.429-438.
Igolkina, A.A., et al, 2025. *Nat. Genet.*, 57, pp.2289-2301.
Jaegle, B., et al, 2023. *Gen. Biol.*, *24*(1), p.44.
Lian, Q., et al, 2024. *Nat. Genet.*, *56*(5), pp.982-991.
Milesi, P., et al, 2024. *Nat. Comm.*, *15*(1), p.8538.
Salojärvi, et al, 2017. *Nat. Genet.*, *49*(6), pp.904-912.
Weissensteiner, et al, 2020. *Nat. Comm.*, *11*(1), p.3403.
Wood, D.P., et al, 2025. *bioRxiv*, pp.2025-07.
Zhou, et al, 2022. *Nature*, *606*(7914), pp.527-534.

**2. Work plan** – MAX 3 PAGES

*Project team* In addition to myself as the appointed researcher and project leader (**James Ord**, Academy of Finland Research Fellow), the project is supported by leading expertise in plant genetics / genomics at the Centre for Excellence in Tree Biology (TreeBio): **Jarkko Salojärvi**, who notably led the original birch genome project (Salojärvi et al 2017), and **Tanja Pyhäjärvi** who is expert in population genetics and genomic breeding of trees and, as a member of the GenTree consortium will also assist in accessing samples from two GenTree field sites (see Fig. 2). Also contributing from TreeBio are **Ville Koistinen** (doctoral candidate in Pyhäjärvi group and expert in GWAS) and **Wenbo Luo** (doctoral candidate in Salojärvi group and expert in high molecular weight DNA isolation). For the pangenome graph component, we are supported by excellent expertise in pangenomics from the Plant Health and Adaptation Group at Kew Gardens, UK (**Daniel Wood**, Future Leaders Fellow), who have successfully generated a pangenome for Ash (*Fraxinus*) (Wood et al, 2025). The pangenome will be greatly enhanced by the inclusion of material from near the range edges. For this, we are assisted by **Dragos Postolache** (National Institute for Research and Development in Forestry, Romania) to obtain specimens growing in the Transylvania region of Romania and from Moldova, and **Filipe Costa e Silva** (Universidade de Lisboa Instituto Superior de Agronomia) to obtain samples from Northern Portugal representing the westernmost edge of the European range. The project is therefore a truly pan-European effort.

*Sample collection* Sampling will be carried out in late winter / early spring (March to April), as young leaves become ready to emerge. Bud burst is generally later for more northern accessions but can be induced by placing but branches in water. In any case, fresh young leaves will be taken and frozen in liquid nitrogen. Most of the individuals to be sequenced will be sampled from the UH greenhouse or surrounding fields. For Lithuanian, Spanish, Romanian, Moldovan, and Portuguese samples, live branches with unburst buds or young leaves will be shipped via express courier and young leaves snap frozen in liquid nitrogen upon arrival at UH.

**Fig. 4.** Representative gel image of three plant HMW DNA samples extracted in our lab. We typically obtain bands that are above the longest standard in the ladder, around 40kb. Photo: Wenbo Luo.

*DNA extraction and sequencing* At our lab at the University of Helsinki, **Wenbo Luo** has worked to optimise the technique for extraction of high molecular weight (HMW) DNA from birch samples, with fresh young leaves giving best results (**Fig. 4**). The ability to purify long DNA fragments allows us to take full advantage of long read sequencing technology and assemble very high-quality genomes. For example, extraction of HMW DNA from shrub birch (*B. humilis*) using this method and subsequent sequencing on the PacBio Revio platform resulted in the assembly of an entire chromosome as a single contig (i.e. basic assembly without the need for additional scaffolding). The HMW DNA samples will be sequenced at BIDGEN (UH sequencing facility), which routinely carries out long read sequencing and has a proven track record of high-quality service. The PacBio Revio flow cells currently used at the facility deliver 80-100Gb of data. Approx. 30x

coverage is typically recommended for genome assembly, which for *B. pendula* (440Mb genome) would require approx. 13Gb of data. Sequencing 16 libraries across three flow cells would deliver 15-19Gb of data per sample.

*Genome assembly and pangenome graph construction* Computational work required for assembling the genomes, pangenome, and downstream analysis will be supported by Centre for Scientific Computing (CSC) which provides free high performance computing infrastructure for academic researchers in Finland. Long reads from each individual will be assembled into contigs using a well-established genome assembler such as hifiasm which, for diploid birch, typically recovers two sets of haplotype sequences each in the region of 400-440Mb. As birch is quite heterozygous, each haplotype assembly will be considered as a distinct individual in the pangenome. Thus, with 18 individual specimens sequenced, the pangenome will be comprised of 36 haplotypes. Two versions of the pangenome graph will be generated and compared: (1) an 'all-against-all' alignment with pggb which has the potential to reveal complex structural variants (e.g. nested inversions), and (2) a reference-based approach whereby all sequences from all individuals are aligned to one high-quality reference genome with minimap2 followed by SV calling with Sniffles2 which will potential uncover more SVs but less complex ones. Individual haplotype genomes will be annotated for gene models using, for example BRAKER3 or HELIXER. The annotation pipeline will be run multiple times for each genome, taking only genes that are identified repeatedly in a given genome. This step is important as it will allow us to more robustly identify accessory genes – those that are actually present in some individuals but not others. Although RNAseq is not budgeted in the current proposal, from the same samples it will also be possible to extract RNA to later obtain transcriptomic evidence for specimen-specific gene models, thus increasing confidence particularly in putative accessory genes.

*Downstream analyses* To determine the 'completeness' of the pangenome, we will compute a saturation curve of structural variants by counting the number of SVs contributed by each subsequent individual added to the pangenome in a random order. With the predicted protein sequences of all predicted genes across all genomes, Orthofinder will be used to identify homologues between sequenced individuals, allowing us to identify the 'core' genes (present in all individuals) and 'accessory' genes (present in only some). Like the SV saturation curve, counting the numbers of genes contributed by each subsequent individual added to the pangenome in a random order will give an indication of the completeness of the gene repertoire of the pangenome. Using functional annotation software such as PANNZER2, we will see examine the putative functions of accessory genes, especially those unique to range edges. To test the utility of the pangenome graph as a reference for short read-based analyses, we will first align short reads from natural populations (WGS reads from 170 individuals and BSseq reads from 25 individuals) to the pangenome and compare alignment rates and mapping quality metrics to those derived from alignment to the linear reference genome. In the case of WGS, reads will be aligned to the pangenome with vg giraffe and to the linear reference genome with BWA-mem, while BSeq reads will be aligned to the pangenome with methylGrapher and to the linear reference genome with Bismark. The natural population samples derive from many of the same locations covered by the planned pangenome. Secondly, a GWAS analysis will be carried out by **Ville Koistinen** using short read WGS data from 410 *B. pendula* samples from Eastern Finland.

Phenotypic data from over 7000 offspring have been generated including growth rates and various morphological traits, which are used to calculate breeding values for each of the 410 sequenced parental trees. Short reads from the 410 individuals will be aligned to the pangenome with the vg giraffe aligner, the output of which can be converted to a conventional VCF file of genotyped SVs that can then be used for GWAS using conventional tools designed for SNP-based GWAS (e.g. PLINK or GEMMA).

*Timeline and reporting* Initial payment will cover shipment of samples from collaborators and extraction of high molecular weight DNA. In the **interim technical report** we will detail the final sample collection and the quality of the extracted DNA (quantity, fragment lengths, purity). After approval of the interim technical report, we will submit one third of the samples for sequencing at BIDGEN on one PacBio Revio flow cell. Upon receipt of the PacBio HiFi data we will assess the quality of the initial data and genome assemblies. We will also test the pipeline for generating a draft pangenome graph using the assemblies generated up to that point. Details of data, genome assembly, and draft pangenome will be included in the **final technical report**. After approval of the final technical report, we will proceed with sequencing the remaining samples on the other two PacBio flow cells, then proceed to generate the remaining genome assemblies, the final pangenome graph, and downstream analyses. **The genome assemblies and pangenome graph comprise the major deliverables of the project which we anticipate to be completed within one year of the project's commencement**. The downstream analyses and preparing of a manuscript detailing the results may take an additional year.

*Plan for data accessibility and publicisation* The genome assemblies and the PacBio HiFi read data used to generate them will be uploaded to the **European Nucleotide Archive no later than one year after commencement of the project**, and the pangenome graph and other associated files will be uploaded to a Zenodo archive, to be freely accessible to the community. To protect the novelty of the EFI-funded work and the impact of its publicisation, an embargo on free access to the data will be imposed for up to one year to allow time to prepare a pre-print detailing the results, which will be made available on BioRxiv prior to submission to a journal. Code used to generate the genome assemblies, pangenome graph, and downstream analyses will be made available on GitHub. Finally, we will apply to present the results of the project at the EVOLTREE conference in 2027.

*Cost breakdown* The requested EVOLTREE Opportunity grant is to cover the core project costs of DNA extraction and sequencing (sequencing provided by BIDGEN sequencing facility at UH), totalling €9,948: Sixteen PacBio libraries priced at €300 each = €4,800, three PacBio flow cell runs priced at €1,600 each = €4,800, and four Monarch HMW DNA Extraction Kits (New England Biolabs) priced at €87 each = €384. The small residual will contribute to any additional reagent costs and shipping costs which will otherwise be covered by other funding.

*Risks and mitigation* A great advantage of the proposal is that the majority of specimens to be sequenced are available on site, while only a small proportion of the planned specimens require collection and shipment from outside Finland. If, due to logistical challenges, it is not possible to obtain a viable sample from a given location, we will seek alternative samples of similar origin, for example from the University Botanic Gardens.