A Strategy for Studying Epigenetic Diversity in Natural Populations: Proof of Concept in Poplar and Oak

Isabelle Lesur Kupin, Odile Rogier, Mamadou Dia Sow, Christophe Boury, Alexandre Duplan, Abel Garnier, Abdeljalil Senhaji-Rachik, Peter Civan, Josquin Daron, Alain Delaunay, et al.

A Strategy for Studying Epigenetic Diversity in Natural Populations:

Proof of Concept in Poplar and Oak

***Lesur I.*[*12], Rogier O.*[*3], Sow M-D. *[*45], Boury C.*[1], Duplan A. [35], Garnier A. [6], Senhaji-Rachik A.*[13], *Civan P.*[4], *Daron J.*[7], *Delaunay A.*[5], *Duvaux L.* [1], *Benoit V.*[3], *Guichoux E.* [1], *Le Provost G.* [1], *Sanou E.* [8], *Ambroise C.* [8], *Plomion C.* [1], *Salse J.* [4], *Segura V.* [39], *Tost J.* [6], *Maury S.* [25]

1 INRAE, Univ. Bordeaux, BIOGECO, F-33610 Cestas, France.

2 HelixVenture, F-33700 Mérignac, France.

3 INRAE, ONF, BioForA, F-45075 Orléans, France.

4 INRAE/UCA UMR GDEC 1095. 5 Chemin de Beaulieu, F-63100 Clermont Ferrand, France.

5 LBLGC, INRAE, Université d'Orleans, EA 1207 USC 1328, F-45067 Orleans, France.

6 Centre National de Recherche en Génomique Humaine, CEA-Institut de Biologie François Jacob, Université Paris-Saclay, F-91000 Evry, France.

7 Institut Pasteur, Université Paris Cité, CNRS UMR2000, Insect-Virus Interactions Unit, F-75724 Paris, France.

8 LaMME, 23 Bd. de France, F-91037 Évry Cedex, France.

9 UMR AGAP Institut, Univ Montpellier, CIRAD, INRAE, Institut Agro Montpellier, F-34398 Montpellier, France.

* These authors contributed equally to this work

↩ Authors for correspondence: stephane.maury@univ-orleans.fr

Lesur I. ***isabelle.lesurkupin@inrae.fr***

Rogier O. ***odile.rogier@inrae.fr***

Boury C. ***christophe.boury@inrae.fr***

Sow M-D. ***mamadou-dia.sow@inrae.fr***

Duplan A. ***alexandre.duplan@univ-orleans.fr***

Garnier A. ***agarnier@cnrgh.fr***

Senhaji-Rachik A. *abdeljalil.senhajirachik@aphp.fr*

Civan P. *peter.civan@inrae.fr*

Daron J. *Josquin.daron@pasteur.fr*

Delaunay A. *alain.delaunay@univ-orleans.fr*

Duvaux L. *ludovic.duvaux@inrae.fr*

Benoit V. *vanina.benoit@inrae.fr*

Guichoux E. *erwan.guichoux@inrae.fr*

Le Provost G. *gregoire.le-provost@inrae.fr*

Sanou E. *doedmond.sanou@univ-evry.fr*

Ambroise C. *christophe.ambroise@univ-evry.fr*

Plomion C. *christophe.plomion@inrae.fr*

Salse J. *jerome.salse@inrae.fr*

Segura V. *vincent.segura@inrae.fr*

Tost J. *tost@cnrgh.fr*

Maury S. *stephane.maury@univ-orleans.fr*

Highlight:

We developed a strategy and a workflow for quantifying epigenetic diversity in natural populations combining whole genome and targeted capture sequencing for DNA methylation.

Abstract

These last 20 years, several techniques have been developed for quantifying DNA methylation, the most studied epigenetic marks in eukaryotes, including the gold standard method, whole-genome bisulphite sequencing (WGBS). WGBS quantifies genome-wide DNA methylation but has several inconveniences rendering it less suitable for population-scale epigenetic studies. The high cost of deep sequencing and the large amounts of data generated prompted us to seek an alternative approach. Restricting studies to parts of the genome would be a satisfactory alternative had there not been a major limitation: the need to select upstream targets corresponding to differentially methylated regions (DMRs) as targets. Given the need to study large numbers of samples, we propose a strategy for investigating DNA methylation variation in natural populations, considering the structural complexity of the genomes with their size and their content in unique as coding regions versus repeated regions as transposable elements. We first identified regions of highly variable DNA methylation in a representative subset of genotypes representative of the biological diversity in the population by WGBS. We then analysed the variations of DNA methylation in these targeted regions at the population level by Sequencing Capture Bisulphite (SeqCapBis). The entire strategy was then validated by applying it to another species. Our strategy was developed as a proof of concept on natural populations of two forest species: *Populus nigra* and *Quercus petraea*.

Keywords and Abbreviations:

DNA Methylation, Epigenetics, Epigenomics, Methylome, Natural population, Oak, Poplar, Transposon Insertion Polymorphism, SeqCapBis, WGS, WGBS

WGS: Whole-genome sequencing

WGBS: Whole-genome bisulphite sequencing

SeqCapBis: Sequencing capture bisulphite

DMC: Differentially methylated cytosine

DMR: Differentially methylated region

SMP: Single methylation polymorphism

PCA: Principal Component Analysis

TIPs: Transposable element insertion polymorphisms

TSD: Target site duplication

Introduction

It is becoming increasingly evident that epigenetic processes both influence phenotype and interact with genetic variation (Bossdorf *et al.*, 2008; Rajpal *et al.*, 2022; Gallusci *et al.*, 2023), but questions remain about the possible ultimate control of epigenetic variation by genetic variation. Indeed, laboratory studies on plants and animals have shed light on some of the general features of epigenetics, with important evolutionary implications (Chapelle and Silvestre, 2022; Husby, 2022; Rajpal *et al.*, 2022). However, studies aiming to determine the relative contributions of genetic and epigenetic variation to phenotypic variation in natural populations can provide important information relative to this debate, particularly if performed in wild populations encountering naturally different levels of environmental complexity, with different genetic structures and dynamics, and natural ecological processes. For example, recent work on the Olympia oyster showed that both genetic and epigenetic components underlie population divergence in fitness-related traits based on the spatial heterogeneity of abiotic and biotic factors (Silliman *et al.*, 2023). However, additional studies are required, in both plants and animals, to clarify the role of epigenetic variability in wild populations (Tost, 2023).

Epigenetics is defined as modifications that affect gene expression without changing the DNA sequence (Russo *et al.*, 1996); it may partly account for missing heritability (Maher, 2008; Danchin, 2013). The principal molecular support for epigenetic mechanisms in eukaryotic cells is chromatin. Cytosine methylation is one of the commonest epigenetic marks in eukaryotes (Schmitz *et al.*, 2019). In plants, DNA methylation consists in the addition of a methyl group in 5' cytosines and occurs in three contexts: CpG, CHG, and CHH (H = bases A, T, or C) and is maintained at hemi-methylated sequences after replication (Zhang *et al.*, 2018; Lloyd and Lister,

2022) or established *de novo* at unmethylated sequences by the RNA-dependent DNA methylation (RdDM) pathway (Erdmann and Picard, 2020). DNA methylation can be eliminated actively by DNA glycosylases via base excision or passively through a lack of maintenance of methylation after replication. Each cytosine residue may be either methylated or unmethylated, corresponding to a single methylation polymorphism (SMP) by analogy to single nucleotide polymorphism (SNP) for genetic variability. However, methylation analysis is usually performed on cell populations (tissues, organs, entire plants) and methylation rate at a single position is a continuous variable, ranging from 0 to 100% depending on the proportion of methylated cytosines in the cell population. Most epigenomics studies focus on DNA methylation because (1) it can be transmitted through cell division (mitosis and meiosis), (2) it is related to gene expression and to the mobilisation of transposable elements (TE), although relationships are highly complex (Ramakrishnan *et al.*, 2021; Mhiri *et al.*, 2022; Muyle *et al.*, 2022), (3) its dynamics and variation play major roles during development and in response to environmental changes, including stress and priming (Lloyd and Lister, 2022), and (4) it is relatively easy to analyse at the genome scale.

One of the most widely used methods for studying DNA methylation is bisulphite sequencing (Tost, 2022; Singer, 2019), or more precisely whole-genome bisulphite sequencing (WGBS), which provides a complete set of quantitative information about cytosine methylation over the entire genome at single-nucleotide resolution. WGBS is considered as the reference method for genome-wide epigenetic studies (Lister *et al.*, 2009). However, despite the progressive decrease in sequencing costs, epigenetic investigation through WGBS renders this approach expensive, particularly for large genomes or large-scale studies. It also generates large amounts of data requiring an appropriate informatics infrastructure with major computing and storage capacities. These drawbacks may decrease the feasibility of whole-genome sequencing, particularly for population epigenetic studies. As an alternative approach, sequencing costs can

be limited by restricting the study to part of the genome. Approaches targeting a small part of the genome are less demanding in terms of computational requirements because it is easier to reach the required sequencing depth for precise quantification. Such approaches are less expensive than WGBS and therefore more appropriate for studies including many samples. Various methods, including reduced representation bisulphite sequencing (RRBS) (Gu *et al.*, 2011) and EpiGBS (van Gurp *et al.*, 2016), are available. These methods can be used to focus bisulphite sequencing efforts on a small part of the genome, making it possible to achieve an acceptable coverage of the targeted sequences with a limited sequencing effort. However, a major drawback has emerged due to the use of restriction enzymes to fragment DNA. It is not possible to study many genomic regions of interest because they are not located close to a restriction site. The sequencing capture bisulphite (SeqCapBis) (Masser *et al.*, 2016) targeted approach can be used to overcome this drawback. SeqCapBis makes use of hybridisation probes to capture and enrich biologically relevant regions of interest. It is not necessary to have a complete continuous reference genome sequence, a feature of particular interest for studies of non-model forest trees species. Nevertheless, local genome annotation is required for functional interpretation (Gu *et al.*, 2011). Unlike other approaches to studying a portion of the genome based on the use of restriction enzymes, such as EpiGBS and RRBS, the SeqCapBis technique can target most genomic regions of interest, regardless of the presence or absence of restriction sites. Moreover, Buckley *et al*. showed that the level of methylation of human tumours assessed with a targeted sequence capture approach is similar to that obtained by WGBS (Buckley *et al.*, 2022). However, the target regions remain to be defined, particularly for population epigenetic studies.

Considering the pros and cons of these methods, we here develop a strategy for studying the variation of DNA methylation in natural populations. WGBS provides access to the whole methylome but remains expensive, particularly if the number of samples is high. We therefore investigated the ability for capture sequencing (SeqCapBis) to complement the WGBS approach.

We developed an optimised strategy that we validated in three steps (Figure 1): 1. Identification of regions displaying highly variable DNA methylation by WGBS applied to a subset of genotypes from natural populations representative of the biological diversity of these populations; 2. Population-level analysis of the highly variable regions of the optimised reduced epigenome; 3. Validation of the population epigenomics strategy in another species. Overall, we aimed to draw up a roadmap for studying epigenetic diversity in natural populations. Here, as a proof of concept, we studied the variation of DNA methylation at population level by considering two forest tree species commonly studied in population genetics: black poplar (*Populus nigra*) as a model species for the development of our strategy, and sessile oak (*Quercus petraea*) as a test species for validation of our approach. Indeed, most studies of the genetic basis of adaptation in trees have focused on the contribution of standing structural variation to local adaptation (Alberto *et al.*, 2013; Plomion *et al.*, 2018). However, epigenetic mechanisms have recently been explored in forest trees (Sow *et al.*, 2018; Amaral *et al.*, 2020), due to their importance in perennials, in which they can facilitate rapid phenotypic modifications in response to environmental changes. Poplar is a model tree species with a high degree of genetic diversity, fast juvenile growth, a high vegetative propagation capacity, amenability to transformation and an available genome sequence (Tuskan *et al.*, 2006). Over the last decade, poplar has been used to investigate the role of DNA methylation in phenotypic plasticity and adaptation to environmental change (Lafon-Placette *et al.*, 2013, 2018; Zhu *et al.*, 2013; Conde *et al.*, 2017*a*,*b*; Le Gac *et al.*, 2018; Sow *et al.*, 2018, 2021; Vigneaud *et al.*, 2023). Ten natural populations of black poplar (*Populus nigra*), a keystone forest tree of riparian ecosystems, from Western Europe were recently studied to assess the variability of their methylomes and the role of (epi)genetic regulation in driving tree species evolution and adaptation over periods of a few generations (micro-evolution) to several million years (macro-evolution) (Sow *et al.*, 2023). The study concerned was based on WGBS, with the analysis of 20 genotypes from the 10 populations. Here, the analysis of a set of original experimental data is reported and discussed to illustrate our strategy for studying epigenomics in

natural populations of two species. The complete procedure, with bioinformatics and statistical analyses, and the genomics data used in this study are freely available and are broadly applicable to future epigenomic studies of natural populations of plants or animals.

Materials and Methods

A step-by-step bioinformatics manual is available at the public repository protocols.io under DOI dx.doi.org/10.17504/protocols.

Biological samples and genomic DNA extraction

In total, 24 *Populus nigra* cambium and xylem samples were collected from a common garden in Orléans (FRANCE - described in Chateigner *et al.,* 2020) and used here for the capture design and for optimising the experimental conditions for SeqCapbis, respectively (Table 1). The 24 genotypes come from 20 genotypes (two genotypes per population) representative of the geographic range of the species in Western Europe analysed in a previous study (Sow *et al.,* 2023) and four additional genotypes (Loire_SPM-034, Loire_SPM-004, Loire_VDL-052, Ticino_N-30) from the same common garden experiment. The methods used for genomic DNA extraction were described in the previous study (Sow *et al.,* 2023).

Eight *Quercus petraea* bud samples (Tronçais-189, Bercé-193, Grésigne-37, St-Sauvant_6, Besange_82, Lappwald_108, Longchamps_136 and Gohrde-89) were harvested during ecodormancy (*i.e.* early spring) in a common garden experiment located in the North East of France (Sillégny, 48°59'13.4"N 6°07'57.6"E). This common garden contains trees of 103 provenances sampled over the entire distribution area of sessile oak. Among those, we selected

eight individuals corresponding to eight provenances representing the variability of bud burst in this species. Two *Quercus robur* bud samples (Bourran-214 and Bourran-274) from a full-sib progeny were also harvested in a common garden experiment located in South-West France (Bourran, 44°,19',44''N, 0°,21',26''E) (Table1). Indeed, *Q. robur* and *Q. petraea* are congeneric species that live together in the same stands. Thus, interspecific hybridization, leading to greater genetic variability and better local adaptation, is widespread (Leroy *et al.*, 2020). Oak bud samples were placed in liquid nitrogen for storage immediately after collection. DNA was then extracted from two sets of pooled buds with a customised CTAB extraction protocol (Larue *et al.,* 2021). The quantity and quality of the DNA were assessed by spectrometry (NanoDrop™ 8000, Thermo Fisher Scientific, Waltham, MA, USA) and fluorimetry (Qubit™, Thermo Fisher Scientific).

Whole-genome sequencing (WGS) and SNP detection

A preliminary WGS step was required for filtering purposes, to prevent C/T SNPs being interpreted as bisulphite conversions of unmethylated sites (*i.e.* false-positive calls). WGS and SNP calling procedures are fully described elsewhere, together with the data for the 20 poplar individuals studied (Sow *et al.,* 2023) (Figure 1). Briefly, sequencing reads were trimmed with Trimmomatic version 0.38 (Bolger *et al.,* 2014), mapping was performed with BWA mem 0.7.17 (Li, 2013) against the *Populus trichocarpa* V3.1 reference genome and SNPs were detected with three SNP-calling tools: bcftools 1.8 (Danecek *et al.,* 2021), FreeBayes 1.2.0-2 (Garrison and Marth, 2012) and GATK 4.0.11.1 (McKenna *et al.,* 2010). SNPs detected with at least 2 approaches were kept. WGS was performed by the Centre National de Recherche en Génomique Humaine (CNRGH), Institut de Biologie François Jacob, CEA, Evry, France.

The same procedure was used in oak, with the following adjustments: for trimming, an additional cutadapt 1.14 step was added (Martin, 2011) and for mapping, the *Q. robur* reference genome (Haplome V2.3) (Plomion *et al.,* 2018) was used; SNP calling was performed with GATK 3.8 (McKenna *et al.,* 2010) and bcftools 1.6 (Danecek *et al.,* 2021). Computational limitations associated with GATK and FreeBayes due to the very deep sequencing in oak (100X on average) necessitated a reduction of the complexity of each dataset. To reduce redundancy within the WGS dataset, we randomly downsampled sequencing reads over genome regions that are over-covered. We therefore performed a digital normalization step with the KHMER digital normalization method (Crusoe *et al.,* 2015) and reduced the coverage to 30X. The SNP identification step used for C/T filtering is described in Supplementary Fig. S1.

Whole-genome bisulphite sequencing (WGBS)

WGBS has been conducted following Sow *et al.,* 2023, on poplar. The same procedure was applied to the additional DRA-038 sample and the 10 oak samples (Table1). Bisulphite sequencing of the DRA-038 sample, which was performed at the Centre National de Recherche en Génomique Humaine (CNRGH) (Institut de Biologie François Jacob, CEA, Evry, France) on HiSeqX5 Illumina sequencer with 2x150bp chemistry, yielded 216,204,762 read pairs, and 44% of the trimmed reads correctly mapped to the poplar genome. Methylation levels were assessed as previously described (Sow *et al.,* 2023) (Figure 1). SMPs identified with the GALAXY (The Galaxy Community, 2022) pipeline (Dugé de Bernonville *et al.,* 2022; Sow *et al.,* 2023) were then used to identify differentially methylated cytosines (DMCs) and differentially methylated regions (DMRs) with the methylKit R package (Akalin *et al.,* 2012) and custom-developed python scripts (https://doi.org/10.57745/IKNRNM).

WGS and WGBS data for the in-silico detection of transposable element insertion polymorphisms in populations

Fastq sequencing files (poplar WGS and WGBS) were trimmed with the TrimGalore tool (V0.6.5; parameters --q 20 and --paired (Krueger *et al.,* 2023), mapped to the reference genome with BWA (V0.7.17; default parameters; (Li and Durbin, 2009)) for WGS data or with Bismark (V0.19.0; default parameters; due to the EpiTEome requirement (Krueger and Andrews, 2011)) for WGBS data. We then used Samtools (version 1.14; default parameters; (Danecek *et al.,* 2021)) to extract WGS reads from the BAM files not aligned to the reference genome. For WGBS reads, the fasta files for the unmapped reads were generated directly by Bismark with the –un option. These unmapped reads (WGS or WGBS), together with the reference genome and a custom transposable element (TE) dataset, were then processed with TEFLoN (V0.4; (Adrion *et al.,* 2017) (Supplementary Fig. S2) or EpiTEome (Daron and Slotkin, 2017) (Supplementary Fig. S3) to detect *in silico* transposable element insertion polymorphisms (TIPs). TEFLoN searches for TIPs for all proposed TEs in the dataset created from the repeatmasker GFF file and the reference genome with bedtools-GetFastaBed v2.29.2 (Quinlan and Hall, 2010). This dataset is a fastq file containing the identifiers and the superfamily of TEs, with the associated nucleotide sequence. Using a file containing a list of TE identifiers, the EpiTEome tool searches for a specific batch of TEs in a TE dataset created with a Python algorithm (https://doi.org/10.57745/IKNRNM) from the repeatmasker GFF file (https://data.jgi.doe.gov/refine-download/phytozome?genome_id=533). This file contains information such as TE IDs, chromosomes, start and stop positions, family and superfamily. For TEFLoN, the following parameters were used: quality mapping --q 20, data type --dt pooled, standard deviation --sd 20, and 3 reads in at least one sample, --n1 = 3, with 3 reads

summed across all samples --n2 = 3; whereas the EpiTEome parameter used was --l 100 to specify read size. We searched for TEs with a minimal length of 1,000 bp.

Identification of target regions for the SeqCapBis design

The 20 *Populus nigra* genotypes (Sow *et al.,* 2023) were used to identify the regions to be targeted in the SeqCapBis approach. Each methylation context in each species was considered separately and the data were processed in R (v3.5.1) ('R Core Team (2021)'). First, forward and reverse strands were merged for the CG context only. SMP matrices in each context (CG, CHG and CHH) were then generated. SMPs colocalizing with either a TE or a C/T SNP (see the WGS and SNP detection section) were removed. We retained only positions with a minimum coverage of 7X per sample and we tolerated 30% missing data. We used custom-developed Python scripts (https://doi.org/10.57745/IKNRNM) to combine all samples for each methylation context into a single matrix and to quantify methylation in 1 kb sliding windows over the entire reference genome in 250 bp steps. Each window, with its associated methylation level, was considered as a potential candidate region for targeting in the SeqCapBis capture design. However, we focused on regions displaying high levels of differential methylation between populations. We performed a three-step analysis procedure to identify these regions. We first identified 1 kb windows corresponding to differentially methylated regions in each methylation context with two strategies. Strategy I (STANDARD DEVIATION OF THE MEANS) involved calculating mean C-methylation levels by averaging the methylation level across all cytosines in each window for each individual (C/Cov). We then calculated the standard deviation of this mean across individuals. Strategy II (MEAN OF THE STANDARD DEVIATIONS) involved calculating the standard deviation of methylation between individuals for each cytosine residue. We then calculated the mean standard deviation

for all the cytosine residues in a window. We defined the threshold for outlier detection as (Q3+1.5*(Q3-Q1)) for poplar in the three methylation contexts, Q1 and Q3 being the first and the third quartiles, respectively. Sequencing depth in oak was, on average, about twice that in poplar. We therefore considered a more stringent threshold for the CHH context in oak, using the threshold (Q3+3*(Q3-Q1)). Windows with a variance above this threshold corresponded to the retained DMRs. Finally, only the overlapping windows retained in both strategies I and II were taken into account for merging (bedtools-2.27.1) and removing sequence redundancy between methylation contexts. For oak, the identified DMRs in the CHH methylation context exceeded the maximum size of the capture design. We therefore selected 12,000 CHH windows at random.

Optimization of targeted methylated sequence Capture (SeqCapBis)

A different design was used for each species: a set of 120 bp probes was selected to capture 18 Mb of each genome (Agilent, https://earray.chem.agilent.com/suredesign/). The targeted regions corresponded to the regions identified as differentially methylated between populations. Custom targeted genome bisulphite sequencing was performed with SureSelect XT Methyl-Seq Target Enrichment (Agilent, Santa Clara, CA, USA) according to the manufacturer's recommendations. We assessed the impact of five experimental variables probe dilution (1:1, 1:8, 1:10 and 1:16), amount of input DNA (500 ng, 600 ng, 750 ng, 935 ng, 1000 ng and 3000 ng), DNA fragmentation technique (acoustic *vs.* enzymatic shearing) and number of PCR cycles (14 or 15 cycles) on the DRA-038_CC sample and a set of four degraded DNA samples corresponding to another four poplar individuals (Loire_SPM-034, Loire_SPM-004, Loire_VDL-052, Ticino_N-30). In total, 18 sets of experimental conditions were assessed by comparing SeqCapBis and WGBS (Table 2). The corresponding libraries were fragmented into fragments of about 200 base pairs, which were then subjected to end repair, A-tailing and the ligation of methylated adaptors before hybridisation

with custom probes for 16 h. Bisulphite conversion was then performed with the EZ-DNA Methylation-Gold Kit (Zymo Research, Irvine, CA). A first PCR was performed with eight or nine cycles, followed by a 2nd PCR with six cycles for indexing. Final libraries were quantified by fluorescence with the Quant-iT dsDNA High Sensitivity Assay Kit (Thermo Fisher Scientific) and pooled to an equimolar concentration. The size of the pool was assessed on a TapeStation 4200 system (Agilent Technologies) and its concentration was estimated by qPCR on a LC480 II system (Roche, Basel, Switzerland) using the QIAseq Library Quant kit (Qiagen, Hilden, Germany). All the capture experiments were performed at the PGTB (doi:10.15454/1.5572396583599417E12) and sequencing was performed at the GeT-PlaGe facility on an Illumina NovaSeq 6000, using 2x150bp chemistry.

DNA methylation was assessed in a five-step process: i) read trimming and quality control, ii) read mapping onto the reference genome, iii) duplicate removal, iv) identification of the methylated Cs (mCs) in all sequence contexts, v) extraction of the mCs in each context (Figure 1). Read trimming and quality control were performed with Trimgalore 0.6.5, FastQC 0.11.9 and MultiQC 1.9. Reads with a Phred score above 20 were retained and the Illumina adaptors were removed. The trimmed reads were then mapped against the *Populus trichocarpa* V4.1 genome (Tuskan *et al.,* 2006) with BsmapZ 1.1.3 (Xi and Li, 2009; Zynda, 2018). The *stat* and *fixmate* functions of samtools 1.11 were then used to check mapping quality and to fill in mate coordinates and insert size fields, respectively. Duplicates were removed with the *markdup* function of samtools 1.11 and methylation level were assessed with the *Methratio.py* script from BsmapZ. In each methylation context, sequencing depth was filtered at 10X with the Methylkit package of R v1.18.0 (Akalin *et al.,* 2012). Finally, we set up a bash script (*splitting.sh)* to obtain methylation files for each sample in the three contexts (CG, CHG and CHH). Consistency between the SeqCapBis and WGBS results was assessed by performing a correlation analysis (calculation of Pearson coefficients) on the methylation data (mC positions) for each context with the methylkit package

of R V1.26 (Akalin *et al.,* 2012). Only methylated positions with a sequencing depth of at least 10X common to both SeqCapBis and WGBS were considered. We then normalized read coverage, using the median to calculate the scaling factor. Principal component analysis (PCA), heatmap clustering, DMCs and MA plots were also run on the methylation matrices to assess the consistency between SeqCapBis and WGBS results.

Regarding oak, SeCapBis was performed on four samples (Bezange_82 *i.e.* B82, Berce_193 *i.e.* B193, St Sauvan_6 *i.e.* S6 and Troncais_189 *i.e.* T189) to quantify the similarity of methylation profiles between SeqCapBis and WGBS approaches. The trimmed reads were mapped onto the *Quercus robur* haplome V2.3 (Plomion *et al.*, 2018). As for poplar, we targeted 18 Mb of the genome based on the identified DMRs between the 10 provenances (representative of the variability of bud burst in this species). Samples were multiplexed at equimolar concentrations in a single pool and were jointly subjected to paired-end mode sequencing (2 x 150 bp) in one lane of an Illumina NovaSeq6000 S4 flow cell. The same bisulphite sequencing procedure and bioinformatics pipeline as for poplar were applied to oak, with the following adjustments: 1 µg of input DNA, 1:8 probe dilution and 14 PCR cycles. These adjustments were established following the optimisation of the approach in poplar.

Results

Whole-genome (bisulphite) sequencing to identify differentially methylated regions for the capture design.

We used WGBS data for a collection of 20 poplar genotypes from 10 natural populations (two per population) representative of the geographic range of the species in western Europe (Sow *et al.,* 2023) to identify genomic regions displaying high levels of variability for DNA methylation in natural populations for use as target regions for the SeqCapBis approach (Figure 1). Only positions

common to all genotypes were considered in an initial matrix, with up to 30% missing data tolerated, resulting in a matrix corresponding to 5,077,664 positions in CG, 14,740,512 in CHG and 80,951,501 in CHH contexts (Table 3). We then minimised the bias on methylation calls from bisulphite treatment by filtering the corresponding matrices in the three contexts for SNPs based on the WGS data (Supplementary Table S1), to prevent C/T SNPs being interpreted as a bisulphite conversion of unmethylated sites (*i.e.* false-positive calls) (Figure 1). After filtration, 4,671,065, 14,010,527 and 78,127,449 SMPs were retained in the CG, CHG and CHH contexts, respectively. We also fixed a threshold of 7X coverage, to minimise the rate of methylation call errors. This led to the STR-10 genotype being discarded because it did not reach the minimum level of coverage required. Transposable element (TE) sequences are difficult to analyse in SeqCapBis. Indeed, mishybridisation can be avoided only if the probes used are designed to bind to genomic sequences that are repeated only infrequently. Integrated TEs are also repeated frequently within the genome and were therefore also filtered out (Figure 1, Table 3). The remaining positions were then used to create 1 kb sliding windows at 250 bp intervals with *Methylkit* software. The number of windows identified with this approach was 1,413,389 for the CG context, 1,389,938 for the CHG context and 1,463,413 for the CHH context (Table 3). These windows were used to identify the regions displaying the strongest differential methylation between *P. nigra* populations according to two strategies of variance calculation (see Methods section; Figure 1). Strategy I identified 45,663 windows in the CG context, 82,835 in the CHG context and 88,675 in the CHH context, whereas strategy II identified 26,555, 15,702 and 4,986 regions in the CG, CHG and CHH contexts, respectively (Table 3). Finally, only genomic regions identified by both methods in each of the methylation contexts, corresponding to 9.85 Mb, 7.62 Mb and 3.15 Mb in the CG, CHG and CHH contexts, respectively, were considered for use in SeqCapBis design (Figure 1). This corresponded to a total of 17.84 Mb of non-redundant target regions for SeqCapBis.

Test using whole-genome (bisulphite) sequencing data to detect in-silico *TE insertion polymorphisms in populations to complete the capture design.*

TE sequences were filtered for the design of specific probes for the SeqCapBis approach, but WGS and WGBS data can nevertheless provide useful information about TE insertion polymorphisms (TIPs) between populations (Domínguez *et al.,* 2020). These TIPs correspond to candidate loci displaying variable DNA methylation between populations that could be used to complete the capture design with bordering regions of detected TIPs (Figure 1). Two tools for identifying TIPs were tested: TEFLoN 0.4 (Adrion *et al.,* 2017) using the WGS data and EpiTEome 1.0 (Daron and Slotkin, 2017) using the WGBS data. Both tools are based on the principle of "soft-clipped" unmapped reads (Yan *et al*., 2021). These unmapped reads are cut into two fragments of variable size (k-mer) that are mapped onto the reference genome and the TE library (Supplementary Fig. S4).

TEFLoN was tested on unmapped WGS reads from three poplar populations: Val d'Allier (ALL-14 and ALL-19 genotypes), Dranse (DRA-038 and DRA-045) and Paglia (PG-31 and PG-34). The custom TE library contained 23,728 TEs from 12 different superfamilies (Supplementary Fig. S5). For each population, we retained only the filtered sites predicted for both individuals. We obtained 683 TIPs for Paglia, 747 for Dranse and 823 for Val d'Allier (Supplementary Fig. S6). Only 29 TIPs were common to all three populations (Supplementary Fig. S6) and only 282 TEs were involved in the TIPs identified in the three populations (Figure 2a, Supplementary Fig. S6). An evaluation of the proportions of TE families predicted by TEFloN revealed an enrichment in helitrons and a depletion of gypsy elements. Most of the predicted TIPs had limited numbers of

insertion sites, except for Helitron-N2 and Gypsy-71, which had 195 and 184 insertion sites, respectively, in the Val d'Allier population (Figure 2a). We characterised the TIPs identified by TEFLoN based on the corresponding methylome data obtained by WGBS (Sow *et al.,* 2023) and published Mobilome-seq data for poplar (*P. tremula x alba) (Sow et al.,* 2021) An analysis of the methylome profiles of the TIPs predicted by TEFLoN revealed that more than 70% of TEs were methylated in at least one of the three methylation contexts considered (CG and CHG > 25% and CHH > 10%). However, a similar proportion was found for all poplar TE sequences in the methylome data, suggesting that there is no specific methylation signature for TIPs predicted by TEFLoN (Supplementary Fig. S7). Similarly, the poplar Mobilome-seq database (*P. tremula x alba*) corresponding to 4,828 active TEs (74 families, classified according to depth-of-coverage values) revealed no differences in activity between the TE families predicted by TEFLoN and the total set of active TE families (Supplementary Fig. S8).

We then tested the only available tool, EpiTEome, using WGBS reads (Daron and Slotkin, 2017) . However, EpiTEome was unable to generate data within a reasonable time frame with the available computing resources (Sow *et al.,* 2023), the complete collection of TEs and the whole genome sequence of poplar. A subset of 2,427 TEs, including TEFLoN-predicted TEs such as the Gypsy-23, Gypsy-27 and Gypsy-71 families, was tested for one genotype from the Paglia population (PG-31). A copy of Gypsy-27 and a new inserted TE (TIP) were detected at positions 3,808,778 - 3,808,933 on chromosome 13 with a target site duplication (TSD sequence) at 3,808,811 (Figure 2b) with 5′ and 3′ flanking split reads overlapping at the TSD sequence generated by the TE insertion. EpiTEome also revealed methylation percentages of 100% (CG), 6% (CHG) and 3% (CHH) for the original TE copy and 100% (CG), 66% (CHG) and ~2% (CHH) for the newly inserted TE copy, suggesting specific CHG hypermethylation. However, further analyses will be required before any firm conclusions about a general methylation signature for TIPs can be drawn (Figure 2c). This preliminary test for TIPs combining WGS and WGBS data

provides proof of concept and is promising for the evaluation of structural and epigenetic polymorphisms of transposable elements between natural populations (Figure 1).

Optimisation of the targeted bisulphite capture approach

A custom design for capture bisulphite sequencing was developed based on our identified DMRs in poplar (Figure 1). In total, 17.84 Mb of sequence corresponding to the 25,434 DMRs was covered by 339,658 probes. Capture bisulphite sequencing was applied to the DRA-038 sample (named 'DRA-038_CC' from the Dranse black poplar population, Tables 1 and 2) to determine the optimal conditions for probe dilution, amount of input DNA, DNA fragmentation type (acoustic shearing *vs.* enzyme) and number of PCR cycles (Table 2). Sequencing generated between 22,546,694 (E08) and 36,867,205 (E27) reads (Supplementary Table S1). The percentage of reads aligning with the poplar reference genome for each sample ranged from 43% (E05) to 48.8% (E30), corresponding to the expected values for a duplicated genome (Nunn *et al.,* 2021). PCR duplication levels ranged from 27.4% (E01) to 44.5% (E10). The total number of mCs ranged from 16,183,391 (E30) to 20,354,557 (E13). The percentage of reads on target ranged from 51% (E30) to 63.9% (E01), with a mean value of 58.71% (Supplementary Table S1).

We investigated whether experimental conditions could affect methylation scores by comparing SMPs between SeqCapBis samples by correlation analyses. Correlation coefficients ranged from 0.87 to 0.99 in the CG context, from 0.89 to 0.99 in the CHG context and from 0.82 to 0.95 in the CHH context (Supplementary Fig. S9, Supplementary Fig. S10, Supplementary Fig. S11). All SeqCapBis samples from DRA-038_CC displayed very high correlations despite highly contrasted experimental conditions: from 0.96 to 0.99 for the CG context, from 0.95 to 0.98 for the CHG context and from 0.85 to 0.92 for the CHH context (Table 4). Principal component

analysis (PCA) revealed that most of the variance between samples was explained by the sample type and type of fragmentation (accoustic *vs.* enzymatic shearing) (Figure 3a, Supplementary Fig. S12). This structure also emerged from the heatmap (Figure 3b, Supplementary Fig. S12), which showed that samples from the Dranse genotype (DRA-038_CC) fragmented by Covaris formed one cluster separate from the other poplar samples and those fragmented enzymatically. Based on these results, we decided to focus on DRA-038_CC samples for the comparison between WGBS and SeqCapBis results. Different amounts of input DNA (500, 600, 750, 1000 and 3000 ng) and probe dilutions (1, 1:8, 1:10 and 1:16) were tested for SeqCapBis (Table 2). PCA revealed differences between SeqCapBis and WGBS (PC2 axis) results but also differences between the samples for SeqCapBis, particularly for the CHH context, in which lower DNA input (750, 600 and 500 ng) was associated with greater variability (Figure 3c, Supplementary Fig. S12).

We investigated the effect of probe dilution on methylation status by comparing undiluted probe (dilution 1) with dilutions of 1:8 and 1:10 (Table 5). To identify differentially methylated cytosines (DMCs), we set the methylation threshold at 25% and the minimum cut-off for outlier detection at a p-value below 0.05 using Benjamini-Hochberg correction (FDR).

When dilutions 1 and 1/8 were compared, only 216, 544 and 78 cytosines were identified as differentially methylated cytosines (DMCs) among the 315,168, 875,228 and 4,272,393 cytosine residues in the CG, CHG and CHH contexts, respectively (Figure 3d). Similarly, 267/327,546, 622/890,652 and 105/4,265,543 DMCs were found in the CG, CHG and CHH contexts for comparisons between dilutions 1 and 1/10 and representing between 0.002 and 0.08 % of the total cytosine analysed (Table 5). We retained the following options for further analyses with the SeqCapBis assay: input DNA = 1000 ng, fragmentation by acoustic shearing (Covaris) and a 1:8 probe dilution.

We then compared WGBS and SeqCapBis findings by considering only the cytosine positions identified by both methods, *i.e.* 180,686 positions, in the 3 contexts (CG, CHG and CHH;

Table 5). Correlation coefficients ranged from 0.83 to 0.99 showing high similarity among data (Table 4, poplar data). To estimate the bias of using SeqCapBis instead of WGBS to determine DNA methylation variations in populations, we calculate how many of the 180,686 positions will be predict as differentially methylated cytosine (DMC) using a standard approach with the methylkit package (Akalin *et al.,* 2012; Sow *et al.* 2021) according to our previously published criteria in poplar (p-value>0,05 and methylation differential >25%; Sow et al., 2021; see Table 5). For probe dilution 1 the comparison between SeqCapBis and WGBS data showed that the percentages of significant DMC ranged from 1.54 to 2.33 % in the 3 contexts (Table 5). These results showed that WGBS and SeqCapBis give similar results for methylation status.

Validation of the population epigenomics strategy with oak species

For validation of the proposed strategy with the optimal experimental conditions established in poplar, we realized on oak samples the same first steps using WGS and WGBS data to obtain the SeqCapBis design (Figure 1) and then test it by comparing our optimized SeqCapBis with WGBS data on the same oak samples. We used a collection of 10 oak samples corresponding to eight *Quercus petraea* belonging to eight provenances representing the variability of bud burst in this species and two *Quercus robur* genotypes, a congeneric species that lives together in the same stands and hybridizes with *Q. petraea*. On average, WGBS analysis generated 301,429,759 read pairs, 59.35% of the trimmed reads were correctly mapped to the reference genome and mean sequencing depth was 56.19X (Supplementary Table S2). The same filtering steps were applied as for poplar, based on SNP data, TE positions and a 7X coverage threshold. Following filtering, 4,256,014 cytosines were identified in the CG context, 14,429,546 in the CHG context and 96,137,559 in the CHH context (Supplementary Table S2). The remaining methylated

positions were grouped into 2,174,494 CG, 2,163,372 CHG and 2,201,370 CHH windows. The selection of the most differentially methylated regions (DMRs) in oak individuals by strategies I and II led to the identification of 35,122 CG regions (*i.e.* 47,616,052 bp), 33,710 CHG regions (*i.e.* 47,117,872 bp) and 58,518 CHH regions (*i.e.* 81,002,085 bp). The union of these regions resulted in 93,019 DMRs corresponding to 142 Mb of sequence. As we intended to capture only 18 Mb of the oak genome for bisulphite sequencing, we selected a subset of 796 DMRs (*i.e.* 989,704 bp), 3,276 DMRs (*i.e.* 4,037,192 bp) and 14,135 DMRs (*i.e.* 19,638,387 bp) in the CG, CHG and CHH contexts, respectively. These DMRs were merged, resulting in 14,435 DMRs, corresponding to 19,638,387 bp of sequence to be targeted. As 19,638,387 bp exceeds the maximum sequence length allowed by the Agilent SeqCapBis design (18 Mb), we applied more stringent filtering criteria to reduce the targeted sequence to 16,147,346 bp. Finally, a set of 140,249 probes (120 bp) was designed by Agilent to cover 99.58% of these DMRs.

To compare SeqCapbis and WGBS data in oak, four of our *Q. petraea* samples were used (Bezange_82, Berce_193, St Sauvan_6 and Troncais_189) (Table 1). Following our optimized capture bisulphite sequencing conditions (input DNA = 1000 ng, fragmentation by acoustic shearing (Covaris) and a 1:8 probe dilution), we obtained a mean of 62,391,030 reads per sample, with about 21.97% PCR duplicated reads. The 47.94% of reads correctly mapped onto the oak reference genome included 24.70% on-target reads. Sequencing depth exceeded 10X for 67.89% of the target sequences (Supplementary Table S3). We identified 110,957 positions present in all four individuals and common to WGBS and SeqCapBis in the CG context. The correlation coefficients between WGBS and SeqCapBis for methylation at the same positions ranged from 0.94 for B193 and S6 to 0.95 for T189 and B82 in the CG context (Table 4). By considering the samples separately, we were able to increase the mean number of shared positions to 182,573 per sample. Correlation coefficients ranged from 0.93 for T189, B193 and S6 to 0.94 for B82 in the CG context (Table 4). Regarding the CHG context, correlation coefficients ranged from 0.93

(T189, B193, S6) to 0.94 (B82) while it ranged from 0.72 (B193, S6) to 0.81 (B82, T189) for the CHH context (Table 4). When comparing WGBS and SeqCapBis methylated cytosines, only 1.26%, 1.72%, 1.82% and 1.40% of the Cytosines were identified to be DMCs in the CG context for B82, T189, B193 and S6, respectively. In the CHG context, these figures dropped to 0.54%, 0.67%, 1.06% and 0.91% for B82, T189, B193 and S6, respectively. Finally, only 0.04%, 0.04%, 0.41% and 0.66% of the Cytosines were identified to be differentially methylated cytosines (DMCs) in the CHH context for B82, T189, B193 and S6, respectively (Table 5). In all three methylation contexts, we showed that the methylation level of most Cytosines is similar with both WGBS and SeqCapBis technologies. These findings successfully validate our strategy for studying DNA methylation variation in natural populations.

Discussion

While there is strong evidence to suggest that epigenetic variability plays a role in phenotypic diversity, which can occur within a generation, potentially enabling rapid responses to environmental changes (Rajpal *et al.*, 2022; Gallusci *et al.*, 2023), the relative contributions of genetic and epigenetic variations to phenotypic variability remain unclear. Studies on wild populations could shed some light on this aspect by estimating how epigenetic variation may be affected by both environmental and genetic variations (Chapelle and Silvestre, 2022; Husby, 2022; Rajpal *et al.*, 2022).

Due to technical limitations, such as the lack of a reference genome sequence, size and complexity of genomes, sequencing costs and the need for downstream bioinformatic analysis, the methylome has been investigated at population level in only a limited number of plant and animal species (Chapelle and Silvestre, 2022; Husby, 2022; Rajpal *et al.*, 2022). For example,

Hagmann *et al.* (2015) reported an interaction between epigenetics and genetic variability in natural populations of *A. thaliana* in North America, where genetic distance and methylome variation were highly correlated. Environmental conditions led to epigenetic divergence between populations in parallel with a divergence of DNA sequences. The major constraints limiting studies of epigenetic diversity at the population scale remain the costs of sequencing and data storage requirements. The cost of high-throughput sequencing technologies — such as the gold standard technique for studying epigenetic variation in plants, WGBS — is falling, but such methods remain very expensive when hundreds or thousands of individuals must be studied or if the species concerned has a large genome. Furthermore, a high-quality reference genome is required for evaluation of the correlation between epigenetics profiles throughout the genome and environmental variation or genetic population structure (Niederhuth *et al.*, 2016).

Studies of epigenetic variation in a reduced representation of the genome (RRBS) rather than the whole genome have been shown to decrease experimental costs and to be a valuable alternative to the WGBS method (Gu *et al.*, 2011; Wang *et al.*, 2015; Trucchi *et al.*, 2016; van Gurp *et al.*, 2016; Paun *et al.*, 2019). The technologies available for targeting part of the genome include EpiGBS (van Gurp *et al.*, 2016), which has been identified as a good candidate method for this purpose (Sepers *et al.*, 2019). Furthermore, as this method does not require a reference genome, it is suitable for use in forest tree species, many of which currently have no established reference genome sequence. However, EpiGBS and RRBS are also subject to several major limitations. Indeed, the main drawback of these technologies is their use of restriction enzymes, necessitating the presence of appropriate restriction sites close to the target regions of interest. We also found significant differences in the results obtained between the enzymatic and acoustic fragmentation methods used for WGBS. We found that the variability between samples could be explained principally by population and the shearing method used, with more marked differences obtained for methods based on restriction enzymes. Furthermore, in comparisons of the WGBS and

EpiGBS results for genomic regions covered by both technologies, methylation levels were found to be very similar for the two techniques in the CG context, but only weakly correlated in the CHG and CHH contexts (van Gurp *et al.*, 2016; Gawehns *et al.*, 2022). In addition, the parts of the genome sequenced depend on the choice of restriction enzyme, limiting the feasibility of comparisons of methylation levels between individuals (Gawehns *et al.*, 2022). As an alternative, part of the genome can be targeted without enzymes in a sequence capture approach known as SeqCapBis. This method uses hybridisation probes rather than enzymatic digestion and can therefore be used to select targets independently of the location of restriction sites. Targeting only part of the genome may decrease costs by decreasing the amount of sequencing required, but it should be borne in mind that bisulphite sequencing is associated with a lower mapping efficiency (Heer *et al.*, 2018). Consequently, larger numbers of reads than for classical targeted sequencing are required for the efficient quantification of methylation levels. A sequencing depth of 8 to 15X is required for correct quantification of the percent methylation at a given locus (Heer *et al.*, 2018). In our study, 24% of sequencing reads mapped on target for oak samples, consistent with the 25.20% correct mapping rate in our previous study on oak with the same technology but without bisulphite treatment (Lesur *et al.*, 2018). However, only 68% of the targets were sequenced with a coverage above 10X. In our previous study (Lesur *et al.*, 2018), a much higher level of target coverage was obtained, with 95.47% of targets covered by more than 10 sequencing reads. This finding confirms previous reports of impaired read mapping following bisulphite treatment. By contrast, Heer et al. (2018) successfully mapped 43% of their bisulphite reads for *Picea abies* onto the exome of this species (Heer *et al.*, 2018), with 72% of their on-target reads presenting a coverage of at least 10X.

Individually, neither WGBS nor SeqCapBis seems to be optimal for studies of population epigenomics. We propose an alternative approach combining the advantages of these two methods: the suitability of SeqCapBis for analyses of hundreds of samples at limited cost, and

the ability of the WGBS approach to ensure the identification of regions throughout the genome displaying differential methylation between natural populations. This approach makes it possible to focus on the variable part of the methylome. We developed a workflow to reduce the costs of population epigenetics studies and we tested our strategy on populations of two tree species widely used for studies of population genetics for local adaptation: poplar and oak (Plomion *et al.*, 2018; Chateigner *et al.*, 2020). This constitutes a significant challenge for forest tree species, given the variability of genome size in these species and the large numbers of repeated sequences they contain. Not all loci are informative; some may have constitutively low or high levels of methylation, regardless of the environmental conditions or genotype (Aliaga *et al.*, 2019), so targeting only the informative variable part of the genome is a more appropriate approach.

Our method involves first performing WGBS on a few individuals (20 of 240 genotypes from 10 natural sites in this study) representative of the natural diversity of the targeted species for the identification of DMRs in the three methylation contexts according to a custom-developed statistical approach to the identification of genomic regions of interest for the SeqCapBis approach. These DMRs best discriminating between populations can then be studied with a SeqCapBis approach applied to large numbers of individuals (hundreds or thousands) representative of the diversity of natural populations. As transposable element (TE) sequences are difficult to analyse in SeqCapBis, we investigated the possible use of WGBS/WGS data for identifying TIPs, which may also correspond to regions of variable DNA methylation at population level that can be relevant for the SeqCapBis design using their bordering regions. TIP detection can be evaluated with real sequencing data and reference genomes (Lerat *et al.*, 2019; Baduel *et al.*, 2021). Here, as a proof of concept, we tested two tools based on the concept of "soft-clipped" reads, EpiTEome and TEFLoN, which use WGBS and WGS data, respectively. TEFLoN had a shorter computation time and a lighter output than EpiTEome for the same allocated computational resources. TEFLoN outputs include several statistics that can be used to evaluate

the proposed predictions with different filter parameters, whereas EpiTEome is more suitable for studies of specific families of TEs and provides a BAM file for visualising the soft-clipped reads mapped onto the genome, the potential TSD and the methylation values of the parental and newly inserted copy of the TE. From a biological perspective, TEFLoN yielded a whole genome estimate of 683 to 823 TIPs detected *in silico*, depending on the population, and representing approximately 2% additional copies relative to the reference genome. Most of these insertions are specific to the population concerned, but we also identified insertion sites common to two or three populations (29 TIPS) specific to metapopulations or to *P. nigra* speciation. Similar experiments on *D. melanogaster* (Adrion *et al.*, 2017) predicted 280 insertion events, with a higher rate of insertion for TE Copia than for other TE superfamilies. TEFLoN can be applied to a whole genome for a full TE library, whereas EpiTEome can test only a limited subset of TEs when genome size is bigger than Arabidopsis one (125 Mb). For example, EpiTEome detected one insertion polymorphism for Gypsy-27 but, for the complete TE collection, it reported only 18 unique TIPs in maize (Daron and Slotkin, 2017) or 11 TIPs in *Arabidopsis* F2 hybrid lines from crosses between wild-type (WT) and elf6-C/ref6-5 (Antunez-Sanchez *et al.*, 2020). Using available methylome and Mobilome-seq data for TEs in poplar (Sow *et al.*, 2021, 2023), we were unable to identify any signature differentiating the TIPs from all TEs. However, the new copy of the TE was hypermethylated relative to the parental site in the CHG context. This suggests that new TE copies may be targeted by the de novo methylation machinery (Mhiri *et al.,* 2022), including the RdDM pathway, leading to their silencing (Fultz *et al.*, 2015). A combination approach may be useful, with potential TIPs detected with TEFLoN and these candidate TIPs then being tested with EpiTEome for further validation of the *de novo* insertion. Molecular biology methods, such as PCR can also be used to validate TIPs, by designing primers binding outside the detected TIPs and within the transposon, based on the physical reads identified with EpiTEome (Antunez-Sanchez *et al.*, 2020). Finally, *in silico* predictions could be improved by optimising the number of individuals per population, optimising populations, and making use of additional tools, such as an

improved version of the SPLITREADER and TEPID pipelines, as previously reported (Baduel *et al.*, 2021). However, the lack of tools for WGBS data (Goerner-Potvin and Bourque, 2018) suggests a need for the development or upgrading of tools such as EpiTEome 1.0. Another promising approach is the use of long-read sequencing technologies, such as HiFi PACBIO or Nanopore sequencing, and dedicated tools, such as LoRTE (Disdero and Filée, 2017).

Long-read sequencing methods have been developed for the quantification of DNA methylation levels over large parts of the genome (Agius *et al.*, 2023), to overcome the limitations of bisulphite sequencing. Indeed, technologies such as Nanopore sequencing from Oxford Nanopore technologies (ONT) and single-molecule real-time sequencing (SMRT) from Pacific BioSciences (PacBio) can prevent the degradation of DNA associated with bisulphite treatment. Furthermore, as no amplification is required, longer native sequences can be mapped onto the reference genome with a higher correct mapping rate (Rand *et al.*, 2017; Gouil and Keniry, 2019). ONT technology has already been successfully used to generate methylation profiles from native tumour DNA in humans (Euskirchen *et al.*, 2017) and, more recently, Schall *et al.* (2023) generated genome-wide profiles by Nanopore sequencing in domestic dogs and showed concordance with WGBS and RRBS data. However, the actual cost and efficiency of long-read sequencing for methylome analysis have yet to be established for large-scale sampling, such as that required for population studies.

Here, with our two-step method combining WGBS, SeqCapBis and the corresponding bioinformatics and statistics workflow, the regions of the entire genome most variable for DNA methylation between populations can be scanned at population level. Such analysis has been extended to the European populations of black poplar (*Populus nigra*) and sessile oak (*Quercus petraea*) with over 240 genotypes per species (data not shown). Such data will help to better understand the potential role of epigenetic in local adaptation of trees.

Conclusion

Variations of DNA methylation at population level are of considerable interest in the context of the adaptation of long-lived organisms to rapid climate change. Here, we provide experimental data validating and demonstrating the applicability of our strategy in three steps: i) the validation of SeqCapBis data relative to the gold standard WGBS method, ii) the technical improvement of a reliable SeqCapBis approach to reduce costs (acoustic DNA shearing, 1:8 probe dilution, DNA quality, 1 µg of input DNA per sample, and pooling for an Illumina NovaSeq6000 S4 flow-cell) and iii) validation of the entire strategy established on the basis of poplar in another tree species (oak) using distinct tissues or organs (cambium and xylem in poplar versus buds for oak) to further study epigenetic during wood formation or bud phenology, respectively, according to previous works (Chateigner *et al*., 2020; Le Provost *et al*., 2023; Sow *et al*., 2023).

Our workflow, including bioinformatics, statistics, and the corresponding genomics data, is freely available and broadly applicable to plants and animals in studies of natural population, and more broadly large scale sampling epigenomics. Relative to the WGBS gold standard method, the combination of WGBS with SeqCapBis seems to be a relevant alternative making it possible to decrease the associated costs and bioinformatic analyses required considerably.

Acknowledgements

This work was funded by the ANR EPITREE (ANR-17-CE32-0009-01). We thank the Genotoul bioinformatics platform Toulouse Occitanie (Bioinfo Genotoul, https://doi.org/10.15454/1.5572369328961167E12) for providing computing and storage resources. WGS and WGBS sequencing were performed at the Centre National de Recherche en Génomique Humaine (CNRGH), Institut de Biologie François Jacob, CEA, Evry, France. All the capture experiments were performed at the PGTB (doi:10.15454/1.5572396583599417E12). SeqCapBis Sequencing was performed at the GeT-PlaGe core facility, INRAe Toulouse.

Conflict of interest

The authors have no competing interests to declare.

Funding

The main source of funding was the ANR through the project EPITREE (ANR: ANR-17-CE32-0009-01).

Data Accessibility

Raw sequencing data are available from the NCBI – SRA database (https://www.ncbi.nlm.nih.gov/).

The WGS sequencing data for the 10 oak samples studied here are available as BioProject PRJNA818131. WGBS sequencing data for these samples are available as BioProject PRJNA818171. SeqCapbis sequencing data for the four samples are available as BioProject PRJNA918844. WGS sequencing data for the 20 poplar samples studied are available as BioProject PRJNA818172. WGBS sequencing data for the same individuals are available as BioProject PRJNA818172. WGBS sequencing data for the DRA-038 sample used as a reference for determination of the impacts of SeqCapBis experimental conditions on DNA methylation level are available under as BioProject PRJNA913022. Finally, the SeqCapBis sequencing data generated for assessment of the effects of the 18 experimental conditions are available as BioProject PRJNA929323. All intermediate datasets and scripts are available from the Research Data Gouv. INRAe Repository (https://entrepot.recherche.data.gouv.fr/) hosted as a permanent resource by INRAe under doi:10.57745/IKNRNM (https://doi.org/10.57745/IKNRNM). For oak and poplar, SNP datasets and the genomic coordinates targeted in the SeqCapBis experiment are available. Methylation levels in the three contexts (CG, CHG and CHH) with BSmapZ are accessible for WGBS and SeqCapBis. An analysis of *de novo* TE insertions (scripts, input and output files) has also been deposited in this repository. Finally, the script used to split the mCs detected with BSMAPz into the three methylation contexts (CG, CHG, CHH) is available. A step-by-step bioinformatics manual is also available at the public repository protocols.io under DOI dx.doi.org/10.17504/protocols.

References

Adrion JR, Song MJ, Schrider DR, Hahn MW, Schaack S. 2017. Genome-Wide Estimates of Transposable Element Insertion and Deletion Rates in Drosophila Melanogaster. Genome Biology and Evolution **9**, 1329–1340.

Agius DR, Kapazoglou A, Avramidou E, *et al.* 2023. Exploring the crop epigenome: a comparison of DNA methylation profiling techniques. Frontiers in Plant Science **14**, 1181039.

Akalin A, Kormaksson M, Li S, Garrett-Bakelman FE, Figueroa ME, Melnick A, Mason CE. 2012. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. Genome Biology **13**, R87.

Alberto FJ, Aitken SN, Alía R, *et al.* 2013. Potential for evolutionary responses to climate change - evidence from tree populations. Global Change Biology **19**, 1645–1661.

Aliaga B, Bulla I, Mouahid G, Duval D, Grunau C. 2019. Universality of the DNA methylation codes in Eucaryotes. Scientific Reports **9**, 173.

Amaral J, Ribeyre Z, Vigneaud J, Sow MD, Fichot R, Messier C, Pinto G, Nolet P, Maury S. 2020. Advances and Promises of Epigenetics for Forest Trees. Forests **11**, 976.

Antunez-Sanchez J, Naish M, Ramirez-Prado JS, *et al.* 2020. A new role for histone demethylases in the maintenance of plant genome integrity. eLife **9**, e58533.

Baduel P, Quadrana L, Colot V. 2021. Efficient Detection of Transposable Element Insertion Polymorphisms Between Genomes Using Short-Read Sequencing Data. Methods in Molecular Biology (Clifton, N.J.) **2250**, 157–169.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics **30**, 2114–2120.

Bossdorf O, Richards CL, Pigliucci M. 2008. Epigenetics for ecologists. Ecology Letters **11**, 106-115.

Buckley DN, Gooden G, Feng K, Enk J, Salhia B. 2022. Targeted DNA methylation from cell-free DNA using hybridization probe capture. NAR Genomics and Bioinformatics **4**, lqac099.

Chapelle V, Silvestre F. 2022. Population Epigenetics: The Extent of DNA Methylation Variation in Wild Animal Populations. Epigenomes **6**, 31.

Chateigner A, Lesage-Descauses M-C, Rogier O, *et al.* 2020. Gene expression predictions and networks in natural populations supports the omnigenic theory. BMC Genomics **21**, 416.

Conde D, Le Gac A-L, Perales M, Dervinis C, Kirst M, Maury S, González-Melendi P, Allona I. 2017*a*. Chilling-responsive DEMETER-LIKE DNA demethylase mediates in poplar bud break. Plant, Cell & Environment **40**, 2236–2249.

Conde D, Moreno-Cortés A, Dervinis C, Ramos-Sánchez JM, Kirst M, Perales M, González-Melendi P, Allona I. 2017*b*. Overexpression of DEMETER, a DNA demethylase, promotes early apical bud maturation in poplar. Plant, Cell & Environment **40**, 2806–2819.

Crusoe M, Alameldin H, Awad S, *et al.* 2015. The khmer software package: Enabling efficient nucleotide sequence analysis. F1000 Research doi: 10.12688/f1000research.6924.1.

Danchin E. 2013. Avatars of information: towards an inclusive evolutionary synthesis. Trends in Ecology & Evolution **28**, 351–358.

Danecek P, Bonfield JK, Liddle J, *et al.* 2021. Twelve years of SAMtools and BCFtools. GigaScience **10**, giab008.

Daron J, Slotkin RK. 2017. EpiTEome: Simultaneous detection of transposable element insertion sites and their DNA methylation levels. Genome Biology **18**, 91.

Disdero E, Filée J. 2017. LoRTE: Detecting transposon-induced genomic variants using low coverage PacBio long read sequences. Mobile DNA **8**, 5.

Domínguez M, Dugas E, Benchouaia M, Leduque B, Jiménez-Gómez JM, Colot V, Quadrana L. 2020. The impact of transposable elements on tomato diversity. Nature Communications **11**, 4058.

Dugé de Bernonville T, Daviaud C, Chaparro C, Tost J, Maury S. 2022. From Methylome to Integrative Analysis of Tissue Specificity. Methods in Molecular Biology (Clifton, N.J.) **2505**, 223–240.

Erdmann RM, Picard CL. 2020. RNA-directed DNA Methylation. PLoS genetics **16**, e1009034.

Euskirchen P, Bielle F, Labreche K, *et al.* 2017. Same-day genomic and epigenomic diagnosis of brain tumors using real-time nanopore sequencing. Acta Neuropathologica **134**, 691–703.

Fultz D, Choudury SG, Slotkin RK. 2015. Silencing of active transposable elements in plants. Current Opinion in Plant Biology **27**, 67–76.

Gallusci P, Agius DR, Moschou PN, Dobránszki J, Kaiserli E, Martinelli F. 2023. Deep inside the epigenetic memories of stressed plants. Trends in Plant Science **28**, 142–153.

Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing. arXiv doi: 10.48550/arXiv.1207.3907. [Preprint].

Gawehns F, Postuma M, van Antro M, *et al.* 2022. epiGBS2: Improvements and evaluation of highly multiplexed, epiGBS-based reduced representation bisulfite sequencing. Molecular Ecology Resources **22**, 2087–2104.

Goerner-Potvin P, Bourque G. 2018. Computational tools to unmask transposable elements. Nature Reviews Genetics **19**, 688–704.

Gouil Q, Keniry A. 2019. Latest techniques to study DNA methylation. Essays in Biochemistry **63**, 639–648.

Gu H, Smith ZD, Bock C, Boyle P, Gnirke A, Meissner A. 2011. Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. Nature Protocols **6**, 468–481.

van Gurp TP, Wagemaker NCAM, Wouters B, Vergeer P, Ouborg JNJ, Verhoeven KJF. 2016. epiGBS: reference-free reduced representation bisulfite sequencing. Nature Methods **13**, 322–324.

Hagmann J, Becker C, Müller J, *et al.* 2015. Century-scale methylome stability in a recently diverged Arabidopsis thaliana lineage. PLoS genetics **11**, e1004920.

Heer K, Ullrich KK, Hiss M, Liepelt S, Schulze Brüning R, Zhou J, Opgenoorth L, Rensing SA. 2018. Detection of somatic epigenetic variation in Norway spruce via targeted bisulfite sequencing. Ecology and Evolution **8**, 9672–9682.

Husby A. 2022. Wild epigenetics: insights from epigenetic studies on natural populations. Proceedings. Biological Sciences **289**, 20211633.

Krueger F, Andrews SR. 2011. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics **27**, 1571–1572.

Krueger F, James F, Ewels P, Afyounian E, Weinstein M, Schuster-Boeckler B, Hulselmans G, sclamons. 2023. FelixKrueger/TrimGalore: v0.6.10 - add default decompression path.

Lafon-Placette C, Faivre-Rampant P, Delaunay A, Street N, Brignolas F, Maury S. 2013. Methylome of DNase I sensitive chromatin in Populus trichocarpa shoot apical meristematic cells: a simplified approach revealing characteristics of gene-body DNA methylation in open chromatin state. The New Phytologist **197**, 416–430.

Lafon-Placette C, Le Gac A-L, Chauveau D, *et al.* 2018. Changes in the epigenome and transcriptome of the poplar shoot apical meristem in response to water availability affect preferentially hormone pathways. Journal of Experimental Botany **69**, 537–551.

Larue C, Guichoux E, Laurent B, Barreneche T, Robin C, Massot M, Delcamp A, Petit RJ. 2021. Development of highly validated SNP markers for genetic analyses of chestnut species. Conservation Genetics Resources **13**, 383–388.

Le Gac A-L, Lafon-Placette C, Chauveau D, *et al.* 2018. Winter-dormant shoot apical meristem in poplar trees shows environmental epigenetic memory. Journal of Experimental Botany **69**, 4821–4837.

Lerat E, Casacuberta J, Chaparro C, Vieira C. 2019. On the Importance to Acknowledge Transposable Elements in Epigenomic Analyses. Genes **10**, 258.

Leroy T, Louvet JM, Lalanne C, Le Provost G, Labadie K, Aury JM, Delzon S, Plomion C, Kremer A. 2020. Adaptive introgression as a driver of local adaptation to climate in European white oaks. New Phytology **226(4)**, 1171–1182.

Lesur I, Alexandre H, Boury C, Chancerel E, Plomion C, Kremer A. 2018. Development of Target Sequence Capture and Estimation of Genomic Relatedness in a Mixed Oak Stand. Frontiers in Plant Science **9**, 996.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv doi: 10.48550/arXiv.1303.3997. [Preprint].

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics **25**, 1754–1760.

Lister R, Pelizzola M, Dowen RH, *et al.* 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature **462**, 315–322.

Lloyd JPB, Lister R. 2022. Epigenome plasticity in plants. Nature Reviews. Genetics **23**, 55–68.

Maher B. 2008. Personal genomes: The case of the missing heritability. Nature **456**, 18–21.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal **17**, 10–12.

Masser DR, Stanford DR, Hadad N, Giles CB, Wren JD, Sonntag WE, Richardson A, Freeman WM. 2016. Bisulfite oligonucleotide-capture sequencing for targeted base- and strand-specific absolute 5-methylcytosine quantitation. Age **38**, 49.

McKenna A, Hanna M, Banks E, *et al.* 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research **20**, 1297–1303.

Mhiri C, Borges F, Grandbastien M-A. 2022. Specificities and Dynamics of Transposable Elements in Land Plants. Biology **11**, 488.

Muyle AM, Seymour DK, Lv Y, Huettel B, Gaut BS. 2022. Gene Body Methylation in Plants: Mechanisms, Functions, and Important Implications for Understanding Evolutionary Processes. Genome Biology and Evolution **14**, evac038.

Niederhuth CE, Bewick AJ, Ji L, *et al.* 2016. Widespread natural variation of DNA methylation within angiosperms. Genome Biology **17**, 194.

Nunn A, Otto C, Stadler PF, Langenberger D. 2021. Comprehensive benchmarking of software for mapping whole genome bisulfite data: from read alignment to DNA methylation analysis. Briefings in Bioinformatics **22**, bbab021.

Paun O, Verhoeven KJF, Richards CL. 2019. Opportunities and limitations of reduced representation bisulfite sequencing in plant ecological epigenomics. The New Phytologist **221**, 738–742.

Plomion C, Aury J-M, Amselem J, *et al.* 2018. Oak genome reveals facets of long lifespan. Nature Plants **4**, 440–452.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics **26**, 841–842.

R Core Team (2021). https://www.r-project.org/. Accessed October 2023.

Rajpal VR, Rathore P, Mehta S, Wadhwa N, Yadav P, Berry E, Goel S, Bhat V, Raina SN. 2022. Epigenetic variation: A major player in facilitating plant fitness under changing environmental conditions. Frontiers in Cell and Developmental Biology **10**, 1020958.

Ramakrishnan M, Satish L, Kalendar R, Narayanan M, Kandasamy S, Sharma A, Emamverdian A, Wei Q, Zhou M. 2021. The Dynamism of Transposon Methylation for Plant Development and Stress Adaptation. International Journal of Molecular Sciences **22**, 11387.

Rand AC, Jain M, Eizenga JM, Musselman-Brown A, Olsen HE, Akeson M, Paten B. 2017. Mapping DNA Methylation with High Throughput Nanopore Sequencing. Nature methods **14**, 411–413.

Russo VEA, Martienssen RA, Riggs AD. 1996. *Epigenetic mechanisms of gene regulation*. Plainview, N.Y: Cold Spring Harbor Laboratory Press.

Schall PZ, Winkler PA, Petersen-Jones SM, Yuzbasiyan-Gurkan V, Kidd JM. 2023. Genome-wide methylation patterns from canine nanopore assemblies. G3 (Bethesda, Md.), jkad203.

Schmitz J, Güntürkün O, Ocklenburg S. 2019. Building an Asymmetrical Brain: The Molecular Perspective. Frontiers in Psychology **10**, 982.

Sepers B, van den Heuvel K, Lindner M, Viitaniemi H, Husby A, van Oers K. 2019. Avian ecological epigenetics: pitfalls and promises. Journal of Ornithology **160**, 1183–1203.

Silliman K, Spencer LH, White SJ, Roberts SB. 2023. Epigenetic and Genetic Population Structure is Coupled in a Marine Invertebrate. Genome Biology and Evolution **15**, evad013.

Singer BD. 2019. A Practical Guide to the Measurement and Analysis of DNA Methylation. American Journal of Respiratory Cell and Molecular Biology **61**, 417–428.

Sow MD, Allona I, Ambroise C, *et al.* 2018. Chapter Twelve - Epigenetics in Forest Trees: State of the Art and Potential Implications for Breeding and Management in a Context of Climate Change. In: Mirouze M, Bucher E, Gallusci P, eds. Plant Epigenetics Coming of Age for Breeding Applications. Advances in Botanical Research. Academic Press, 387–453.

Sow MD, Le Gac A-L, Fichot R, *et al.* 2021. RNAi suppression of DNA methylation affects the drought stress response and genome integrity in transgenic poplar. New Phytologist **232**, 80–97.

Sow MD, Rogier O, Lesur I, *et al.* 2023. Epigenetic Variation in Tree Evolution: a case study in black poplar (Populus nigra). bioRxiv.

The Galaxy Community. 2022. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. Nucleic Acids Research **50**, W345–W351.

Tost J. 2022. Current and Emerging Technologies for the Analysis of the Genome-Wide and Locus-Specific DNA Methylation Patterns. Adv Exp Med Biol., **1389**, 395-469.

Tost J. 2023. Do not be scared of the genome's 5th base – Explaining phenotypic variability and evolutionary dynamics through DNA methylation analysis. Mol. Ecol. Resour., **23**, 1473-1476.

Trucchi E, Mazzarella AB, Gilfillan GD, Lorenzo MT, Schönswetter P, Paun O. 2016. BsRADseq: screening DNA methylation in natural populations of non-model species. Molecular Ecology **25**, 1697–1713.

Tuskan GA, DiFazio S, Jansson S, *et al.* 2006. The Genome of Black Cottonwood, Populus trichocarpa (Torr. & Gray). Science **313**, 1596–1604.

Vigneaud J, Kohler A, Sow MD, *et al.* 2023. DNA hypomethylation of the host tree impairs interaction with mutualistic ectomycorrhizal fungus. The New Phytologist **238**, 2561–2577.

Wang K, Li X, Dong S, Liang J, Mao F, Zeng C, Wu H, Wu J, Cai W, Sun ZS. 2015. Q-RRBS: a quantitative reduced representation bisulfite sequencing method for single-cell methylome analyses. Epigenetics **10**, 775–783.

Xi Y, Li W. 2009. BSMAP: whole genome bisulfite sequence MAPping program. BMC Bioinformatics 10, 232.

Yan C, He J, Luo J, Wang J, Zhang G and Luo H. 2021. SIns: A Novel Insertion Detection Approach Based on Soft-Clipped Reads. Front. Genet. **12**, 665812.

Zhang H, Lang Z, Zhu J-K. 2018. Dynamics and function of DNA methylation in plants. Nature Reviews Molecular Cell Biology **19**, 489–506.

Zhu R, Shevchenko O, Ma C, Maury S, Freitag M, Strauss SH. 2013. Poplars with a PtDDM1-RNAi transgene have reduced DNA methylation and show aberrant post-dormancy morphology. Planta **237**, 1483–1493.

Zynda G. 2018. BSMAPz. https://github.com/zyndagj/BSMAPz

Tables

TABLE 1:

List of *P. nigra* and oak samples (*Q. petraea* and *Q. robur*) associated with each sequencing technique. Eight *Q. petraea* and two *Q. robur* were sequenced by WGS and WGBS. Among the *Q. petraea* individuals, four individuals were also sequenced by SeqCapBis. A total of 24 *P. nigra* samples were considered in this study: 20 individuals were sequenced by WGS, 21 individuals were sequenced by WGBS and 5 individuals were sequenced by SeqCapBis.

	WGS	WGBS	SeqCapBis
POPLAR			
SAMN27655113 Rhin_STR-010_13	1	1	0
SAMN27655112 Ticino_SN-7_13	1	1	0
SAMN27655111 Ticino_SN-2_13	1	1	0
SAMN27655110 Rhin_RHN-028_13	1	1	0
SAMN27655109 Paglia_PG-34_13	1	1	0
SAMN27655108 Paglia_PG-31_13	1	1	0
SAMN27655107 Kuhkopf_KUH-44_13	1	1	0
SAMN27655106 Kuhkopf_KUH-36_13	1	1	0
SAMN27655105 Loire_GLY-009_13	1	1	0
SAMN27655104 Loire_GLY-008_13	1	1	0
SAMN27655103 Dranse_DRA-045_13	1	1	0
SAMN27655102 Dranse_DRA-038_13	1	1	0
SAMN32530370 Dranse_DRA-038_CC	0	1	14
SAMN27655101 Basento_BS-37_13	1	1	0
SAMN27655100 Basento_BS-36_13	1	1	0
SAMN27655099 Adour_BDX-003_13	1	1	0
SAMN27655098 Adour_AST-005_13	1	1	0
SAMN27655097 ValAllier_ALL-019_13	1	1	0
SAMN27655096 ValAllier_ALL-014_13	1	1	0
SAMN27655095 Ramieres_1-J31_13	1	1	0
SAMN27654770 Ramieres_1-A26_13	1	1	0
SAMN33219161 Loire_SPM-004	0	0	1
SAMN33219162 Loire_SPM-034	0	0	1
SAMN33219163 Loire_VDL-052	0	0	1
SAMN33219164 Ticino_N-30	0	0	1
OAK			
SAMN26818645 Bezange_82	1	1	1
SAMN26818644 Gresigne_37	1	1	0
SAMN26818643 Lappwald_108	1	1	0
SAMN26818642 Berce_193	1	1	1

SAMN26818641 St Sauvant_6	1	1	1
SAMN26818640 Bourran_274	1	1	0
SAMN26818639 Bourran_214	1	1	0
SAMN26818638 Gohrde_89	1	1	0
SAMN26818637 Troncais_189	1	1	1
SAMN26818636 Longchamps_136	1	1	0

TABLE 2:

Description of the 18 experimental conditions tested in *P. nigra* for the optimization of SeqCapBis.

ID	Sample_ID	Input DNA quantity (ng)	Input DNA quality	DNA Fragmentation	Probe dilution	Number of PCR cycles
E01	Dranse_DRA-038_CC	3000	high-quality	acoustic shearing	1	14
E02	Dranse_DRA-038_CC	1000	high-quality	acoustic shearing	1	14
E03	Dranse_DRA-038_CC	1000	high-quality	acoustic shearing	1/8	14
E04	Dranse_DRA-038_CC	1000	high-quality	acoustic shearing	1/8	14
E05	Dranse_DRA-038_CC	1000	high-quality	enzymatic digestion	1/8	15
E06	Loire_SPM-034	935	degraded	enzymatic digestion	1/8	14
E07	Dranse_DRA-038_CC	500	high-quality	acoustic shearing	1/8	15
E08	Dranse_DRA-038_CC	750	high-quality	acoustic shearing	1/8	15
E09	Loire_SPM-004	1000	degraded	acoustic shearing	1/8	14
E10	Loire_VDL-052	1000	degraded	acoustic shearing	1/8	15
E11	Ticino_N-30	1000	degraded	acoustic shearing	1/8	14
E12	Dranse_DRA-038_CC	1000	high-quality	acoustic shearing	1/10	14
E13	Dranse_DRA-038_CC	1000	high-quality	acoustic shearing	1/16	14
E27	Dranse_DRA-038_CC	1000	high-quality	acoustic shearing	1	14
E28	Dranse_DRA-038_CC	1000	high-quality	acoustic shearing	1/8	14
E29	Dranse_DRA-038_CC	1000	high-quality	acoustic shearing	1/8	14
E30	Dranse_DRA-038_CC	600	high-quality	acoustic shearing	1/8	15
E31	Dranse_DRA-038_CC	1000	high-quality	acoustic shearing	1/10	14

TABLE 3:

Steps in the identification of the targeted sequences for SeqCapBis in *P. nigra*. Once SMPs from raw WGBS were identified, they were filtered (for C/T SNPs, TEs, coverage ≥ 7X) in each methylation context (CG, CHG and CHH). Then, outlier DMRs were identified for the SeqCapBis design with statistic strategies I and II up to 17.84Mb. SMPs, single methylation polymorphisms; SNPs, single-nucleotide polymorphisms; DMRs, differentially methylated regions; TEs, transposable elements.

Steps	CG	CHG	CHH
SMPs detection			
N° of SMPs in samples	8,901,297 – 9,463,906	14,841,207 – 15,597,276	81,561,404 – 86,031,951
N° SMPs in merged matrices (tolerating 30 % NA)	5,077,664	14,740,512	80,951,501
Filtering			
Removal of C SNPs	4,671,065	14,010,527	78,127,449
Removal of TEs positions	4,330,170	13,070,943	72,852,384
coverage (⩾ 7X)	3,267,355	9,498,080	49,019,836
Nber of windows (1kb sliding windows of 250bp)	1,413,389	1,389,938	1,463,413
Statistics			
Outlier DMRs Strategy I	45,663	82,835	88,675
OutlierDMRs Strategy II	26,555	15,702	4,986
OutlierDMRs Strategy I (non-redundant size (Mb))	24.6085	40.8735	47.35225
OutlierDMRs Strategy II (non-redundant size (Mb))	15.31675	8.3515	3.394
OutlierDMRs Strategy I^II overlap (Mb)	9.85125	7.62175	3.15175
Total non-redundant size	**17.84 Mb**		

TABLE 4:

Correlation between SeqCapBis and WGBS samples in *P. nigra* and *Q. petraea*. Pearson correlation coefficients for common SMPs between the WGBS and SeqCapBis samples have been computed for the 14 SeqCapBis experimental conditions tested on the *P. nigra* reference (DRA-038_CC) samples and for the four *Q. petraea* samples in the CG, CHG and CHH contexts.

Experimental condition	Sample	CG	CHG	CHH
P. trichocarpa				
E01	DRA-038_CC	0.99	0.98	0.92
E02	DRA-038_CC	0.99	0.98	0.93
E03	DRA-038_CC	0.99	0.98	0.92
E04	DRA-038_CC	0.99	0.98	0.92
E05	DRA-038_CC	0.96	0.95	0.87
E07	DRA-038_CC	0.99	0.98	0.91
E08	DRA-038_CC	0.99	0.98	0.91
E12	DRA-038_CC	0.99	0.98	0.85
E13	DRA-038_CC	0.99	0.98	0.91
E27	DRA-038_CC	0.99	0.98	0.92
E28	DRA-038_CC	0.99	0.98	0.92
E29	DRA-038_CC	0.99	0.98	0.91
E30	DRA-038_CC	0.98	0.98	0.91
E31	DRA-038_CC	0.99	0.98	0.92
Q. petraea				
WGBS vs. SeqCapBis (1:8)	B82	0.94	0.94	0.81
WGBS vs. SeqCapBis (1:8)	T189	0.93	0.93	0.81
WGBS vs. SeqCapBis (1:8)	B193	0.93	0.93	0.72
WGBS vs. SeqCapBis (1:8)	S6	0.93	0.93	0.72

TABLE 5:

Percentage of differentially methylated cytosines (DMCs) in *P. nigra* and Q. petraea comparing different SeqCapBis probe dilutions or between SeqCapBis and the WGBS samples in the CG, CHG and CHH contexts. To identify DMCs, we set the *methylation threshold at 25% and the minimum cut-off for outlier detection at a p-value below 0.05 using Benjamini-Hochberg correction (FDR).* For *P. nigra,* differentially methylated cytosines (DMCs) between SeqCapbis samples with capture probes diluted to 1, 1:8 or 1:10 were identified in the reference sample DRA-038_CC as well as DMCs between SeqCapbis samples with capture probes diluted to 1:8 and WGBS. For Q. petraea, comparison between dilution 1:8 of SeqCapBis and WGBS was performed in the four samples.

	Sample	CG (%)	CHG (%)	CHH (%)
P. trichocarpa				
SeqCapBis 1:1 *vs.* SeqCapBis 1:8	DRA-038	0.07	0.06	0.002
SeqCapBis 1:1 *vs.* SeqCapBis 1:10	DRA-038	0.08	0.07	0.002
SeqCapBis 1:1 *vs.* WGBS	DRA-038	2.33	2.11	1.54
Q. petraea				
SeqCapBis 1:8 *vs.* WGBS	B82	1.26	0.54	0.04
SeqCapBis 1:8 *vs.* WGBS	T189	1.72	0.67	0.04
SeqCapBis 1:8 *vs.* WGBS	B193	1.82	1.06	0.41
SeqCapBis 1:8 *vs.* WGBS	S6	1.4	0.91	0.66

Figure legends

FIGURE 1: Strategy for population epigenomics combining whole-genome and target genome sequencing. The bioinformatic workflow used in our approach, shows how WGS (red), WGBS (green) and SeqCapBis (blue) analyses are combined for the identification of SMPs in populations. The WGS step (red) is required to remove false C/T SNPs from the SMP dataset. The WGBS step (green), performed on a few individuals representative of the natural diversity of the species, consists in the identification of the SMPs and then DMRs. Afterwards, following a statistical approach, genomic regions of interest (outlier DMRs) are identified and covered by probes. These outier DMRs best discriminating between populations can then be studied with a SeqCapBis approach (blue) applied to large numbers of individuals. Steps 1, 2, 3, 4 and 5 refer to the Methods section. The black star refers to the Agilent web tool for designing a probe set required for the SeqCapBis approach.

FIGURE 2: Detection of Transposon insertion polymorphism (TIP) among 3 black poplar populations**. A.** Distribution of predicted TE families according to their number of insertion sites (1 to 200) detected by TEFLoN in the 3 poplar populations: Val d'Allier (ALL, in pink), Dranse (DRA, in green) and Paglia (PG, in blue). Venn diagram (right corner) of poplar TE families with predicted TIPs using TEFLON shared or not among the 3 populations. **B.** Example of EpiTEome analysis for a new TE Gypsy-27 insertion copy (TIP) in the PG31 genotype (Paglia population) with IGV view of splitted-reads. **C.** Methylation percentages in the three cytosine contexts (CG, CHG and CHH) for the Gypsy-27 parent TE or TIP predicted by EPITEOME.

FIGURE 3: Optimization of the SeqCapBis method with the impact of the experimental conditions tested in the *P. nigra* on the methylation data in the CG context. The Dranse sample corresponds to the *P. nigra* reference sample: Dranse_DRA-038_CC. **A**. Principal component analysis (PCA) on WGBS and

SeqCapBis data for poplar genotypes (Dranse in blue and others in orange, see Table 2) and two DNA fragmentation approaches (Covaris acoustic shearing and enzymatic digestion) for the CG context. **B.** Heatmap (Euclidean distance) based on SeqCapBis and WGBS data for the different experimental setups in the CG context (see Table 2). **C.** PCA on CG methylation data for Dranse samples fragmented by acoustic shearing (Covaris) with five different amounts of input DNA captured with four different probe dilutions. **D.** Volcano plot of differentially methylated cytosines (DMCs) between the SeqCapBis dilution 1:1 and the SeqCapbis dilution 1:8 samples in the CG context.

Figure 1

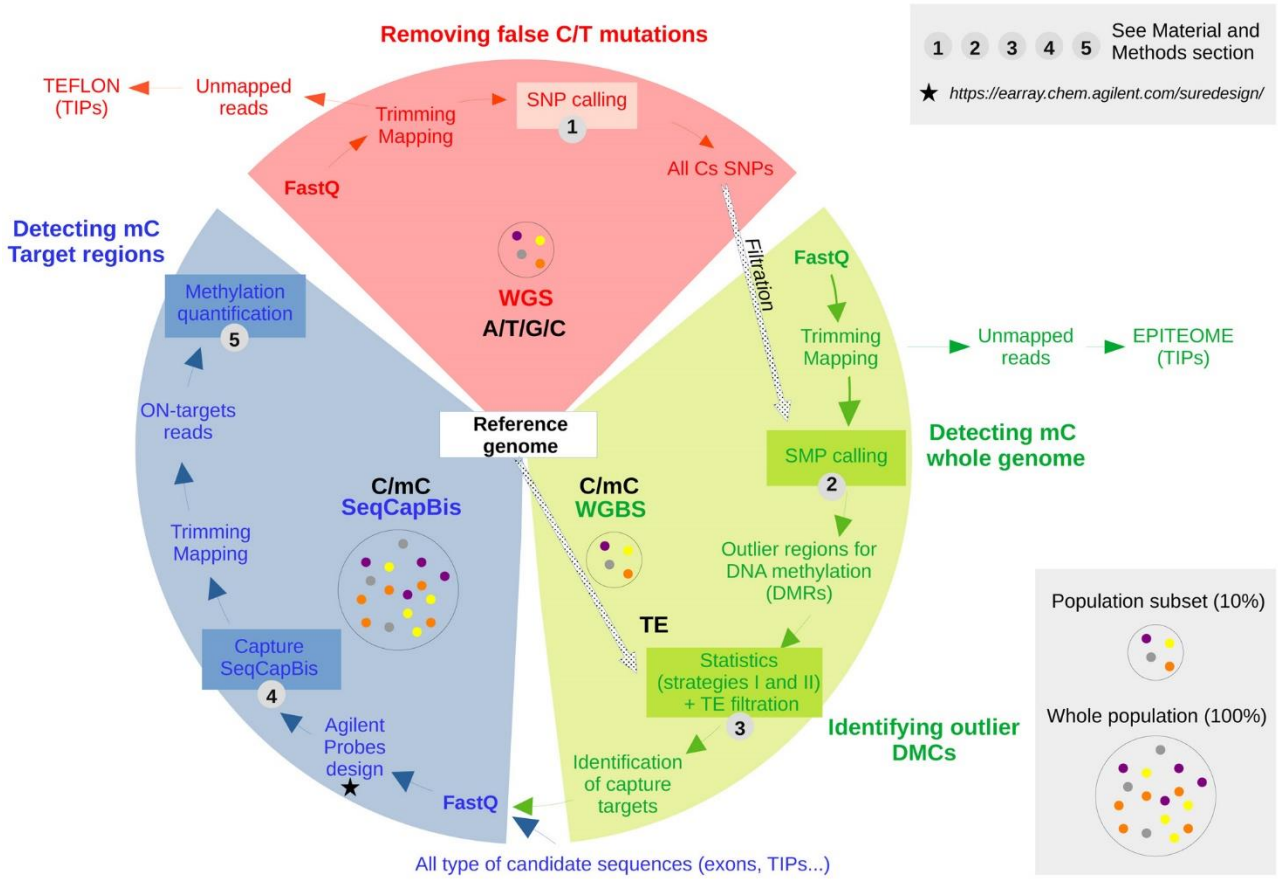

Figure 2

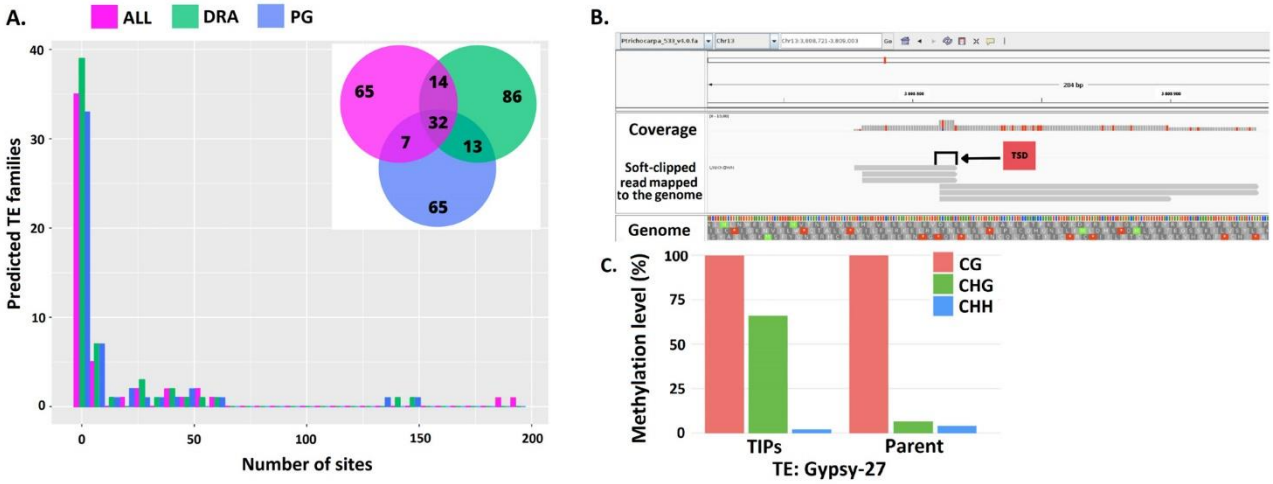

Figure 3