A two-sample tree-based test for hierarchically organized genomic signals

Pierre Neuvial, Nathanaël Randriamihamison, Marie Chavent, Sylvain Foissac, Nathalie Vialaneix

A two-sample tree-based test for hierarchically organized genomic signals

Pierre Neuvial

Institut de Mathématiques de Toulouse, UMR 5219, Université de Toulouse, CNRS UPS, Toulouse, France.

E-mail: pierre.neuvial@math.univ-toulouse.fr

Nathanaël Randriamihamison

Institut de Mathématiques de Toulouse, UMR 5219, Université de Toulouse, CNRS UPS, Toulouse, France

Université de Toulouse, INRAE, UR MIAT, Castanet-Tolosan, France.

Marie Chavent

Université de Bordeaux, CNRS, INRIA, Bordeaux INP IMB, UMR 5251, Talence, France.

Sylvain Foissac

GenPhySE, Université de Toulouse, INRAE, ENVT, Castanet Tolosan, France.

Nathalie Vialaneix

Université de Toulouse, INRAE, UR MIAT, Castanet-Tolosan, France.

E-mail: nathalie.vialaneix@inrae.fr

Summary. This article addresses a common type of data encountered in genomic studies, where a signal along a linear chromosome exhibits a hierarchical organization. We propose a novel framework to assess the significance of dissimilarities between two sets of genomic matrices obtained from distinct biological conditions. Our approach relies on a data representation based on trees. It utilizes tree distances and an aggregation procedure for tests performed at the level of leaf pairs. Numerical experiments demonstrate its statistical validity and its superior accuracy and power compared to alternatives. The method's effectiveness is illustrated using real-world data from GWAS and Hi-C data. Tree distances; Cophenetic distances; Moderated t statistics; p-value aggregation.

1. Introduction

Genetic information is carried by chromosomes and, consequently, is primarily organized along a linear and mono-dimensional genomic axis. Pairwise similarity measures between genomic elements, such as the linkage disequilibrium for Single Nucleotide Polymorphisms (SNPs) or Hi-C interaction matrices for tridimensional chromosome conformation (Dixon et al., 2012), are frequently structured in a specific way. Indeed, the induced (symmetric) similarity matrix presents a pattern of nested diagonal blocks, usually named haplotype groups for linkage disequilibrium and Topologically Associating Domains (TADs) for Hi-C data. Each of these diagonal blocks corresponds to strong correlations between adjacent genomic elements. Tree structures have proved to be an informative representation of such matrices in several genomic contexts (Fraser et al., 2015; Weinreb and Raphael, 2016; Ambroise et al., 2019; Soler-Vila et al., 2020). They have been frequently used to identify structures of interest in the genome, including haplotype groups (Won et al., 2020) and TADs (Soler-Vila et al., 2020). However, their potential for providing a robust summary of the hierarchical organization of the genome for structure comparison remains largely unexplored.

The goal of this paper is to provide a statistical test to compare two datasets with such a natural underlying hierarchical organization. Our approach thus relies on the idea of representing the hierarchical organization of such matrices with trees. Existing methods for statistical assessment on trees and families of trees (Mallows, 1957; Holmes, 2003b) are generally based on bootstrapping an (individual $\times$ descriptor)-matrix to generate a distribution of similarity between elements. This is typically the case in the field of phylogeny for instance (Efron et al., 1996; Holmes, 2003a). As such, these methods cannot be used when the input data are similarities. Other approaches have been designed to compare pairs of trees (Galili, 2015) or pairs of similarity matrices (Fraser et al., 2015). However, by construction, these methods cannot account for within-condition variability, which is essential to assess the significance of differences between conditions.

Here, we introduce a tree-based method to test for significant differences between two sets of similarity matrices obtained from different conditions. Our proposed method,

hereafter referred to as "tree test", proceeds in two steps. First, each similarity matrix is mapped to a multivariate vector encoding the hierarchical structure. This vector corresponds to the cophenetic distances obtained from an adjacency-constrained clustering (Ambroise et al., 2019) of the input similarity matrix. Then, the resulting multivariate two-sample testing problem is addressed by aggregating univariate moderated tests, in order to cope with the typical high-dimensional setting encountered in genomic studies.

We perform extensive numerical experiments based on real GWAS and Hi-C data in order to assess (i) the statistical validity of the approach and (ii) its ability to identify true biological signal. In these experiments, our proposed method is compared to other approaches in order to assess the relevance of the proposed tree-based representation of the input data, and of the chosen statistical test.

The tree test is presented in Section 2. Related works about statistical tests on trees, and multivariate two-samples tests are reviewed and discussed in comparison with our method in Section 3. The results of our numerical experiments are reported in Section 4. This section includes: a simulation study based on GWAS data under the null hypothesis (Section 4.2), an application to a differential analysis of Hi-C experiments (Section 4.3), and a comparison study for simulated biological and technical replicates generated from the same Hi-C data set (Section 4.4). Final remarks are gathered in Section 5.

2. Method

We assume to be given n input $B \times B$ matrices, $(M_i)_{i=1,\dots,n}$ that correspond to a measure of similarity or dissimilarity between genomic elements (genes, genomic intervals, SNPs, ...), with B the number of genomic elements and n the number of observations. Typically, these elements are ordered along the chromosome, *i.e.*, for $j < j' < j''$, the genomic element j' is between the genomic elements j and j'' on the chromosome. As discussed in Ambroise et al. (2019), such matrices frequently present a strong spatial auto-similarity driven by the linear organization of the chromosome. This induces a typical hierarchical structure, which is illustrated in Figure 1 (see also Figure 3 below or Figure 5 (left) in Randriamihamison et al. (2021)).

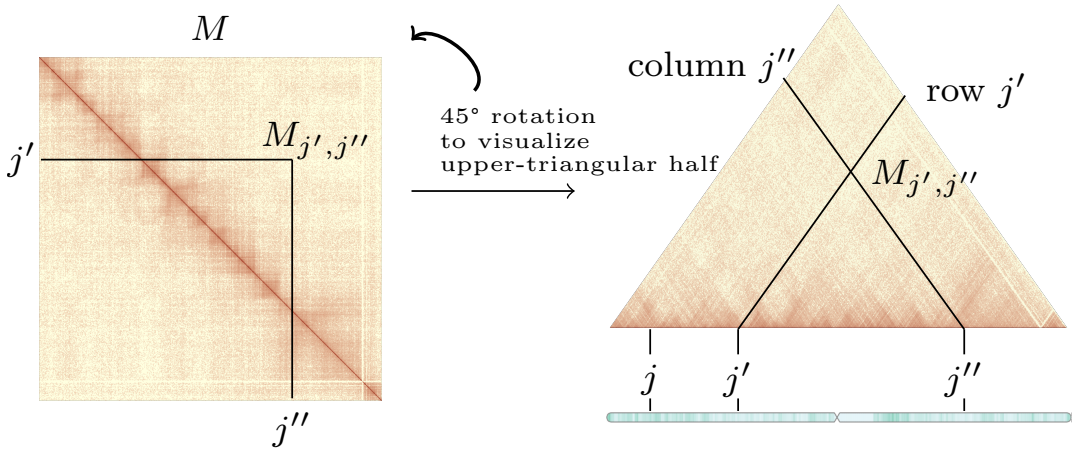

Fig. 1. Genomic similarity matrix obtained from Hi-C data (Dixon et al., 2012). Left: (symmetric) similarity matrix M. The matrix presents a hierarchical structure with nested squares centered on the diagonal. Right: Upper triangular part of the same matrix, with a representation of the linear chromosome underneath. In both panels, elements are ordered according to their chromosomal positions.

The tree test approach proceeds in two steps:

(a) in order to capture relevant information about the hierarchical structure of the input data, we use adjacency-constrained hierarchical clustering (Ambroise et al., 2019) to map each matrix M_i to a dendrogram. This dendrogram is then represented by the vector $\mathbf{X}_i \in \mathbb{R}^p$ (with $p = B(B-1)/2$) of all cophenetic distances between leaf pairs. This construction is described in Section 2.1;

(b) a multivariate two-sample test adapted to the specificity of the data (large p, very small n, strong correlation structure in the multivariate vectors to be compared) is proposed (Section 2.2).

2.1. Mapping similarity matrices to cophenetic distance vectors

In order to capture the hierarchical structure of the matrices $(M_i)_{i=1,\dots,n}$, we perform adjacency-constrained hierarchical clustering with Ward linkage (Ambroise et al., 2019) on each matrix M_i, using the R package **adjclust**. As a result, we obtain a set of

trees $(\tau_i)_{i=1,\dots,n}$ with the same set of leaves $j = 1, \dots, B$ corresponding to the original elements. Note that we use the generic term "similarity" in the sense described in Randriamihamison et al. (2021): it covers kernel-based approaches (*e.g.*, correlation measures) or more general dissimilarity or similarity data for which there exists a valid extension of the above-mentioned hierarchical clustering algorithm.

For each tree τ_i, we then consider the vector $\mathbf{X}_i$ of cophenetic distances between all pairs of leaves (j, j'). The cophenetic distance corresponds to half the length of the shortest path between leaves j and j'. Equivalently, for a dendrogram, this corresponds to the height at which these leaves are merged in the tree, as shown in Figure 2. As the

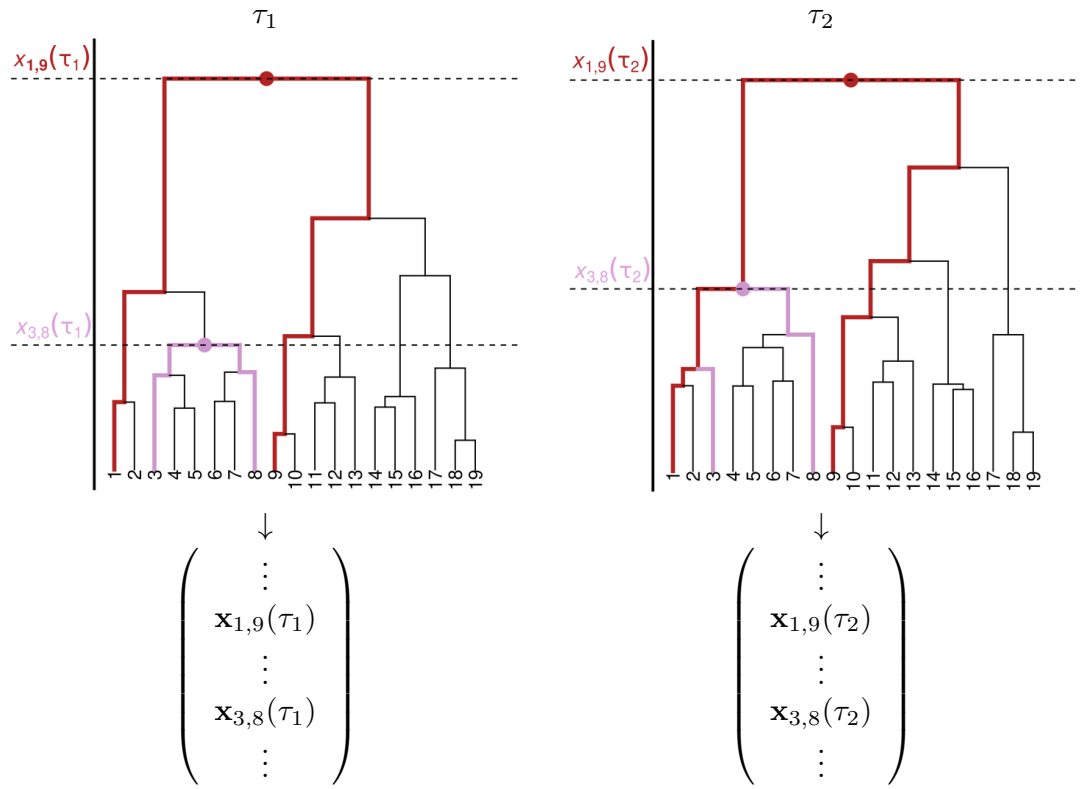

Fig. 2. Mapping from two example trees τ_1 and τ_2, with B leaves each, to their corresponding vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ in $\mathbb{R}^p$, with $p = B(B-1)/2$. The mapping is induced by the cophenetic distances between leaf pairs, which correspond to the height at which leaves are merged. Two examples are highlighted with thicker and colored branches.

number of pairs of leaves, (j, j'), is equal to $p = B(B-1)/2$, this representation maps

the tree τ_i to $\mathbf{X}_i \in \mathbb{R}^p$. This "tree-to-vector" mapping is frequently used to define a distance between trees, as discussed in Section 3.1. The initial matrix/tree comparison problem can thus be rewritten as the search of significant differences between two sets of vectors in $\mathbb{R}^p$.

2.2. Multivariate test for tree structure comparison

Given a sample $(\mathbf{X}_i)_{1 \leq i \leq n}$ of n independent observations, we aim at comparing two conditions denoted by $\mathcal{C}_1$ and $\mathcal{C}_2$, such that $\mathcal{C}_1 \cup \mathcal{C}_2 = \{1, \ldots, n\}$ and $\mathcal{C}_1 \cap \mathcal{C}_2 = \emptyset$. Let $\mu_1 \in \mathbb{R}^p$ and $\mu_2 \in \mathbb{R}^p$ denote the true population means associated to $\mathcal{C}_1$ and $\mathcal{C}_2$, respectively. Our goal is to test the null hypothesis H_0: "$\mu_1 = \mu_2$" against the alternative hypothesis H_1: "$\mu_1 \neq \mu_2$".

To address this p-dimensional testing problem in a genomic context, the following constraints must be taken into account:

(a) usually, genomic experiments are performed with very small sample sizes (typically, $n_1 = n_2 \leq 10$). This makes most permutation or resampling-based approaches inappropriate. Indeed, the number of possible permutations of the two conditions between the n individuals is at most $\binom{n}{n_1}$, which can be too low to obtain sufficient statistical power due to granularity issues, as discussed and illustrated in our numerical experiments;

(b) conversely, the typical number of genomic features considered is larger than 10 (hence, $p \geq 50$), which leads to a high-dimensional problem ($n < p$ or even $n \ll p$) for which, standard parametric multivariate tests (*e.g.*, Hotelling's test; Hotelling (1936)) are ill-defined because they rely on estimating an inverse covariance matrix;

(c) finally, since they are based on a tree structure, the $\mathbf{X}_i$ present a very specific and strong correlation structure, which makes standard alternatives to Hotelling's test not relevant (see Bai and Saranadasa (1996); Chen and Qin (2010); Dong et al. (2016) and further discussion and comparison in Section 3.2).

To handle these constraints, we choose to perform p marginal tests (one for each entry of the cophenetic distance matrix), and summarize the results in a unique test using p-value aggregation. This choice avoids the estimation of an inverse covariance

structure without having to assume independence between tests. Our proposed approach is described below.

2.2.1. Individual statistics

Let $j = 1, \ldots, p$ indicate a leaf pair. We assume that, for $k \in \{1, 2\}$,

$$X_{ij} \sim \mathcal{N}(\mu_{kj}, \sigma_j), \quad i \in \mathcal{C}_k. \tag{1}$$

Noting H_{0j}: "$\mu_{1j} = \mu_{2j}$", we consider the standard two-sample Student test statistic of H_{0j}:

$$t_j := \sqrt{\frac{n_1 n_2}{n}} \times \frac{\overline{\mathbf{X}}_j^{(1)} - \overline{\mathbf{X}}_j^{(2)}}{\hat{\sigma}_j}, \tag{2}$$

where, for $k \in \{1, 2\}$, $n_k = |\mathcal{C}_k|$, $\overline{\mathbf{X}}^{(k)} = \frac{1}{n_k} \sum_{i \in \mathcal{C}_k} \mathbf{X}_i \in \mathbb{R}^p$ and $\hat{\sigma}_j$ is an estimate of σ_j, which is assumed to be identical in the two conditions in (1). Assumption (1) implies that the statistic t_j follows a Student's distribution with $(n-2)$ degrees of freedom (df) under H_{0j} with $\hat{\sigma}_j^2$ defined as the pooled estimate

$$\tilde{\sigma}_j^2 := \frac{1}{n-2} \left(\sum_{i \in \mathcal{C}_1} \left(\mathbf{X}_{ij} - \overline{\mathbf{X}}_j^{(1)} \right)^2 + \sum_{i \in \mathcal{C}_2} \left(\mathbf{X}_{ij} - \overline{\mathbf{X}}_j^{(2)} \right)^2 \right).$$

In the case where many tests are performed with a common ground and a small sample size (not fit for a good estimation of standard deviations), several authors have advocated for a regularization of the estimator $\tilde{\sigma}_j^2$ (Tusher et al., 2001; Smyth, 2004; Tong and Wang, 2012). We thus consider the moderated test statistic t_j defined in Equation (2) associated to a "squeezed" variance estimate $\hat{\sigma}_j^2$ based on a combination of all the empirical estimates $(\tilde{\sigma}_j^2)_{j=1,\ldots,p}$ introduced by Smyth (2004). We refer to Appendix A for a complete and formal definition. Note that this variance moderation step is performed at the level of the whole tested matrix (*e.g.*, considering information coming from $(\tilde{\sigma}_j)_{j=1,\ldots,p}$)†.

Finally, following the original results of Smyth (2004), the p-value for H_{0j} is:

$$\pi_j = 2(1 - F_{\nu_0+n-2}(|t_j|)), \tag{3}$$

†More generally, if several sets of matrices are tested independently (*e.g.*, several chromosomes in the genome), variance moderation can be performed based on the empirical variance estimates of all matrices together.

where F_ν is the cumulative distribution function of the Student's distribution $\mathcal{T}_\nu$ with ν degrees of freedom, and ν_0 corresponds to additional degrees of freedom due to the squeezed variance estimate. In practice, ν_0 is estimated as initially proposed by Smyth (2004) and implemented in the R package **limma** (Ritchie et al., 2015).

2.2.2. Tree-level p-value

Since $\mathrm{H}_0 = \cap_{j=1}^{p} \mathrm{H}_{0j}$, one can test the intersection null hypothesis H_0 by aggregating the corresponding individual p-values $\pi_j, j = 1 \dots p$. Following Lun and Smyth (2014), we use the Simes aggregation method:

$$\pi_{\text{Simes}} := \min \left\{ p\frac{\pi_{(j)}}{j},\ j = 1, \dots, p \right\}, \tag{4}$$

with $\pi_{(j)}$ the jth smallest individual p-value: $\pi_{(1)} \leq \cdots \leq \pi_{(p)}$. This aggregation method produces a valid p-value for the test of H_0 as soon as the Simes (1986) inequality holds. As such, it is valid not only when the p aggregated tests are independent, but also more generally when the p-values associated to true null hypotheses satisfy a technical condition called the Positive Regression Dependence on a Subset (PRDS) condition (Benjamini and Yekutieli, 2001). While proving that the PRDS condition holds for a particular applicative context is challenging, this condition is known as particularly robust to departures of independence (see *e.g.*, Rødland (2006); Goeman and Solari (2014)). In particular, this condition is the weakest known condition under which the widely used Benjamini and Hochberg (1995) procedure controls the False Discovery Rate (Benjamini and Yekutieli, 2001). While we cannot prove theoretically that this condition generally holds for our test, the numerical experiments reported in Section 4 support the empirical validity of the resulting method.

The different steps of the tree test are described in Algorithm 1.

3. Relation to existing literature

In this section, we discuss relations between our tree test and state-of-the-art methods for tree testing, and we compare our test with alternative multivariate testing methods.

Algorithm 1 Tree-based two-sample test for hierarchically organized genomic signals

Require: $(M_i)_{i=1,\dots,n}$ ▷ n similarity matrices

Require: $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2\}$ ▷ partition of observations between conditions

for all $i = 1, \dots, n$ **do** ▷ Map to cophenetic distance vector

  $\tau_i \leftarrow \text{ADJACENCY\_CONSTRAINED\_CLUSTERING}(M_i)$

  $\mathbf{X}_i \leftarrow \text{COPHENETIC\_DISTANCE}(\tau_i)$

end for

for all $j = 1, \dots, p$ **do**

  $\tilde{\sigma}_j^2 \leftarrow \text{POOLED\_VARIANCE\_ESTIMATE}(\mathbf{X}_{\cdot j}, \mathcal{C})$

end for

for all $j = 1, \dots, p$ **do** ▷ Individual tests

  $\hat{\sigma}_j^2 \leftarrow \text{MODERATED\_VARIANCE\_ESTIMATE}((\tilde{\sigma}_{j'}^2)_{j'}, \mathcal{C})$

  $\pi_j \leftarrow \text{MODERATED\_T\_TEST}(\mathbf{X}_{\cdot j}, \hat{\sigma}_j^2, \mathcal{C})$

end for

$\pi_{\text{Simes}} \leftarrow \text{SIMES\_AGGREGATION}((\pi_j)_j)$ ▷ Global test

return π_{Simes}

3.1. Tests on trees

3.1.1. Probabilistic models on tree sets

Existing approaches designed to address the issue of comparing families of trees assume a probabilistic distribution over the space of all possible binary trees with a given set of leaves from which observed objects (here, trees) have been sampled. A statistic with known theoretical distribution is then defined under the assumption that both sets of trees have been sampled from the same initial distribution. Statistical guarantees can then be derived from the comparison between the observed value of this statistic and its expected distribution under the null hypothesis.

This is the path followed by Billera et al. (2001), who assume a probabilistic distribution based on the choice of a distance δ between trees. Elaborating on previous works by Mallows (1957) on ranking models, Billera et al. (2001) consider that the probability of sampling a given tree, τ, is given by

$$f(\tau) := Ke^{-\lambda\delta(\tau,\tau_0)}, \tag{5}$$

where τ_0 is a fixed "central" tree in the set $T(\mathfrak{B})$ of binary rooted trees with B leaves, $1/\lambda$ is a dispersion parameter and K is the normalization constant. Billera et al. (2001) have proposed a constructive iterative approach to compute the central tree τ_0 for their BHV metric.

An alternative to the parametric distribution of Equation (5) is the use of bootstrapping, combined with a convenient tree distance, as proposed by Holmes (2003a). This approach is tailored to cases like phylogenetic trees, where the trees are built from a (observation $\times$ variable)-matrix, and the similarity between observations is calculated from measurements on the variables. Bootstrap samples are then obtained by resampling over the variables, and a nonparametric distribution of the trees is derived from this resampling scheme. By construction, this type of approach cannot handle cases where the input data are similarity matrices that are not obtained from observed measurements on the variables, as is the case for Hi-C data among others. Therefore, such a nonparametric approach is not suited to our case.

3.1.2. Tree distances

Coming back to the parametric probabilistic distribution described in Equation (5), we pinpoint an interesting connection between our method and a specific distance between trees. Our proposed method relies on the representation of a tree τ by the corresponding vector $\mathbf{X}(\tau)$ of cophenetic distances between all of its pairs of leaves, which is described in Section 2.1 and illustrated by Figure 2. This representation defines a mapping from $T(\mathfrak{B})$ to $\mathbb{R}^p$. This mapping is also the basis of a distance between trees introduced by Steel and Penny (1993): the "weighted Path Difference Metric" (wPDM) between two trees τ_1 and τ_2, defined as

$$\text{wPDM}(\tau_1, \tau_2) := 2\|\mathbf{X}(\tau_1) - \mathbf{X}(\tau_2)\|_2,$$

where $\|.\|_2$ stands for the Euclidean distance in $\mathbb{R}^p$. Using the dissimilarity $\delta = \text{wPDM}^2$ in the tree distribution model of Equation (5), the tree distribution density can be rewritten as

$$\forall\, \tau \in T(\mathfrak{B}), \qquad f(\tau) = Ke^{-2\lambda\|\mathbf{X}(\tau)-\mathbf{X}_0\|_2^2},$$

with $\mathbf{X}_0 = \mathbf{X}(\tau_0)$. This probability distribution can also be seen as a distribution for all vectors of $\mathbb{R}^p$ that can be obtained as cophenetic distances for a given tree. Relaxing the latter constraint, it thus defines a distribution over the entire space $\mathbb{R}^p$, which turns out to be the Gaussian distribution $\mathcal{N}_p\left(\mathbf{X}_0, 1/(2\lambda)\mathbb{I}_p\right)$. Hence, our proposal can be interpreted as a test assuming a distribution which is a relaxation of the probabilistic framework of Billera et al. (2001) for wPDM.

This tight connection of the data representation used in our approach to wPDM is central to benefit from other appealing features that wPDM possesses in terms of modeling and computational properties. First, many other distances between trees that have been considered in the literature, in particular edit distances, the Robinson-Foulds distance, the Nearest Neighbor Interchange distance, or the Subtree Pruning and Regrafting distance (Robinson and Foulds, 1981; DasGupta et al., 1997; Bordewich and Semple, 2005) allow to distinguish trees with different topologies (*e.g.*, branching patterns), but they cannot distinguish trees with identical topologies and different branch lengths. However, this latter property is critical to properly compare trees in the numerous applications

where branch lengths are meaningful, including the ones motivating the present paper. To our knowledge, the only existing distances between trees that are able to make this distinction are wPDM and the Billera-Holmes-Vogtmann (BHV) distance (see Holmes (2003b); Chakerian and Holmes (2012) and the R package **distory**). However, wPDM is fast to compute, especially when compared to BHV. Indeed, its computational complexity is of the order of the number p of pairs of leaves, whereas the fastest known algorithm for BHV is of the order of B^4, *i.e.*, $\mathcal{O}(p^2)$, as shown in Owen and Provan (2011).

3.2. Multivariate tests

The test introduced in Section 2.2 can formally be used to compare two sets of multidimensional vectors $(\mathbf{X}_i)_{i\in\mathcal{C}_k}$ for $k \in \{1,2\}$ that do not necessarily come from similarity data. The goal of this section is to compare our test with other approaches that have been proposed to address this multivariate two-sample comparison problem, and to discuss the relevance of the associated model assumptions for the validity of these tests in our genomic context. When $p < n$ and under a Gaussian assumption, the problem of comparing two sets of multidimensional vectors $(\mathbf{X}_i)_{i\in\mathcal{C}_k}$ for $k \in \{1,2\}$ is usually addressed using Hotelling's test statistic (Hotelling, 1936), which is equivalent to the generalized likelihood-ratio test for testing the hypothesis H_0 against H_1. The Hotelling statistic is defined as

$$T^2 = \left(\overline{\mathbf{X}}^{(1)} - \overline{\mathbf{X}}^{(2)}\right)^\top \widehat{\Sigma}^{-1} \left(\overline{\mathbf{X}}^{(1)} - \overline{\mathbf{X}}^{(2)}\right), \tag{6}$$

where $\widehat{\Sigma}$ is an estimator of the covariance matrix Σ between entries of the vectors, assumed to be common to both conditions.

Note that one can interpret T^2 as the squared norm of the decorrelated vector $\left(\overline{\mathbf{X}}^{(1)} - \overline{\mathbf{X}}^{(2)}\right)\widehat{\Sigma}^{-1/2}$. By definition, Hotelling's test requires $p < n$, otherwise $\widehat{\Sigma}$ cannot be invertible. Thus, it is not adapted to the tree framework considered in this paper, which frequently yields high-dimensional situations with $p \gg n$.

Two main directions have been considered to address the issue of the high dimension in Hotelling's tests: either accounting for the dependence between the entries of $\mathbf{X}$ and using a modified decorrelation step as in Shen et al. (2011); Chen et al. (2011), or simply

ignoring this dependence, as in Chen and Qin (2010); Dong et al. (2016). A tradeoff between both solutions has also been proposed in Hébert et al. (2021).

Accounting for dependence generally requires the regularization of the empirical covariance matrix before decorrelation. This is done by replacing $\widehat{\Sigma}$ by $\widehat{\Sigma} + \eta \mathbb{I}_p$, for some $\eta > 0$ large enough so that $\widehat{\Sigma} + \eta \mathbb{I}_p$ is invertible. However, this approach does not come with satisfactory statistical guarantees in high-dimensional situations. Indeed, Chen et al. (2011) obtain a theoretical distribution under the strong hypothesis that n is of the order of p and Shen et al. (2011) use a bootstrapping approach to obtain an estimation of the distribution of their test statistic under H_0, which is also not appropriate for small sample sizes (*e.g.*, $n \leq 10$), as already discussed. This approach is compared to the tree test in the numerical experiments reported below, where it is called "permutation test".

An extreme alternative would be to neglect the dependence between the entries of $\mathbf{X}$. Bai and Saranadasa (1996) derive a statistical procedure where $\widehat{\Sigma}$ is replaced by the identity matrix $\mathbb{I}_p$. This approach has been extended to the case of the large dimension by Chen and Qin (2010). Dong et al. (2016) also assume independence, but account for differences in variances between the entries of $\mathbf{X}$. This leads to a diagonal Hotelling test with $\widehat{\Sigma} = \mathrm{Diag}(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_p^2)$, where $\hat{\sigma}_j^2$ is an estimate of the jth diagonal entry of Σ, the covariance matrix of cophenetic distances. This approach still requires the regularization of the estimation of $(\sigma_j)_{j=1,\ldots,p}$ but is more suited to the large dimension because only p parameters need to be estimated. In this case, Hotelling's statistic can also be seen as a weighted version of wPDM (between the two trees corresponding to averaged cophenetic distance vectors in each group), where the jth entry of the cophenetic distance is weighted by $\hat{\sigma}_j^{-1}$.

However, this independence assumption is unrealistic for cophenetic distance vectors. Indeed, these vectors have many ties, leading to a specific dependence pattern (see an illustration in Section 4.2, Figure 5 right). Ignoring this dependence could compromise the control of type-I error in the test procedure. This approach is compared to the tree test in the numerical experiments reported below, where it is called "diagonal tree Hotelling".

4. Numerical experiments

4.1. Alternative methods

In experiments below, we compare the tree test to several alternatives. First, we have tested different versions of Hotelling's test adapted to large dimensional datasets, following the directions discussed in Section 3.2:

- a diagonal Hotelling's test as Dong et al. (2016). For a fair comparison with the tree test, we also used squeezed variances to implement this test, as described in Section 2.2. We performed this test to compare the cophenetic distance vectors, as done for the tree test, but also to compare the vectors of the original entries from the upper part's matrix. These two versions are named **diagonal tree Hotelling** and **diagonal matrix Hotelling**, respectively;
- a **permutation test** to compare the cophenetic distance vectors. For this test, we performed a standard (regularized) Hotelling test and estimated the p-value from N permutations of the conditions accross samples. This test was implemented using the R package **Hotelling**. Considering the computational burden of this approach and its intrinsic lack of resolution for low sample sizes (see results in the following sections), we only performed it on the matrix entries directly.

Additionally, we also used additional tests designed to work directly on the matrix entries to assess the relevance of the tree model and the data representation described in Section 2.1:

- a **Mantel test** to directly compare matrices. As this test only compares two matrices, we summed the matrices within each condition before using it. Depending on the application, this transformation either makes sense (as for Hi-C data, which are count data, as in Section 4.3) or is probably not relevant (as for LD data, which are correlations, as in Section 4.2);
- the equivalent of the test tree but applied to the entries of the matrices instead of the cophenetic distances. The entries from the upper triangle are tested with moderated Student's tests and results are aggregated with the Simes' procedure. This test is referred to as **matrix test** hereafter.

Table 1. Broad characteristics of the methods considered in the performance evaluation.

Method name	Input data	Testing procedure
tree test	cophenetic	Simes aggregation of moderated t tests
matrix test	similarity	Simes aggregation of moderated t tests
diagonal tree Hotelling	cophenetic	Hotelling test with diagonal covariance
diagonal mat. Hotelling	similarity	Hotelling test with diagonal covariance
permutation test	cophenetic	regularized Hotelling test + permutation
Mantel test	similarity	Mantel test between condition means

A summary of the main characteristics of the different tests considered in experiments is given in Table 1.

4.2. Simulation study under the null hypothesis

To assess the control of the type-I error rate and the distribution of p-values under the null hypothesis, we used real genomic data and randomly split them into two groups. Consequently, no specific signal is expected between the two groups. More precisely, we chose to use GWAS data from the international HapMap project (The International HapMap Consortium, 2003) to perform realistic numerical experiments under the null hypothesis (Figure 3).

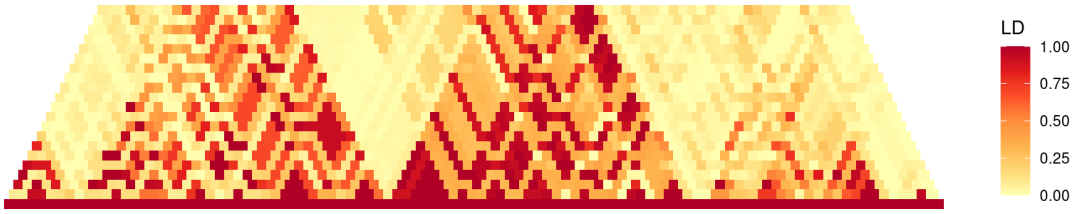

Fig. 3. Linkage disequilibrium for a region of $B = 100$ randomly chosen SNPs (organized along the horizontal axis) from the HapMap project. Only the upper triangular part of the matrix is displayed (to avoid redundancy due to symmetry) and, in addition, it is cut to keep SNP pairs that have no more than 19 SNPs between them.

The GWAS data file contains 603 contiguous SNPs spanning a one megabase region on chromosome 22, in a sample of 90 Europeans. This dataset has been obtained from

the **snpStats** R package. We sampled uniformly at random 60% of the individuals in the initial sample n times to obtain n samples from the same original population. We also randomly selected a region of B contiguous SNPs. Using these samples and the selected SNPs, we built n LD matrices, which are expected to have approximately the same hierarchical structure, because they were obtained from the same population. These matrices were further split into two groups, each with B leaves (the SNPs), on which the tests were performed. Since each split is arbitrary, we expect no true difference between the two groups.

The simulation process was repeated 1,000 times in order to assess the test statistic and p-value distributions. Different variations of the simulation settings (n and B) were also investigated. Simulations were performed using 4 cores and 40Gb of RAM. Figure 4 displays the empirical cumulative distribution function (ECDF) for the p-values obtained over 1,000 simulations, for various values of n and B. The expected behavior is that the ECDF should be close to the diagonal, since we expect Uniform p-values under the null hypothesis. An ECDF below the diagonal indicates that the corresponding test controls the type-I error rate but is conservative and therefore expected to lack power. An ECDF above the diagonal indicates that the corresponding test does not control the type-I error.

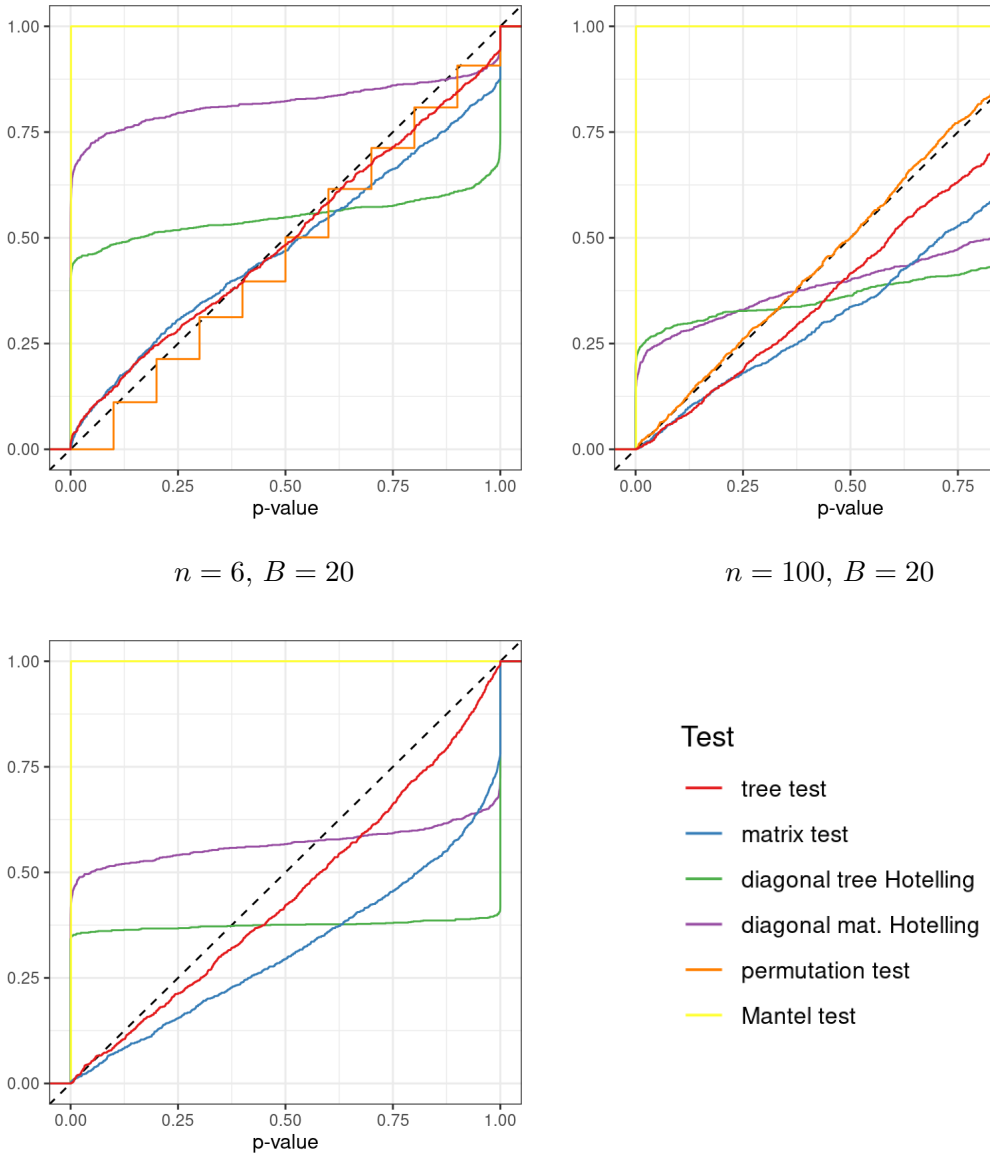

Fig. 4. GWAS: Comparison of the p-value ECDF for the different methods. p-value ECDF for the different methods under a H_0 setting. Each plot corresponds to a tested combination of the parameters n (total number of trees in the two conditions) and B (number of leaves in the trees).

In all settings, the Mantel test and both diagonal Hotelling tests exhibit an excess of false positive results (for a 5% risk, the Mantel test detects 100% of significant differences over the 1,000 simulations for the three settings and the diagonal Hotelling tests detect 24.6-72.4% of significant differences over the 1,000 simulations). For the Mantel test, this bad behavior could be attributed to the fact that the sum (or average) of correlation values is not a relevant operation. However, the inability to deal with replicates within condition is, *per se*, a strong limitation of the Mantel test.

For the diagonal tree Hotelling test, the bad behavior is arguably due to a strong deviation from the diagonal (independent) setting. Among the two versions of the diag-

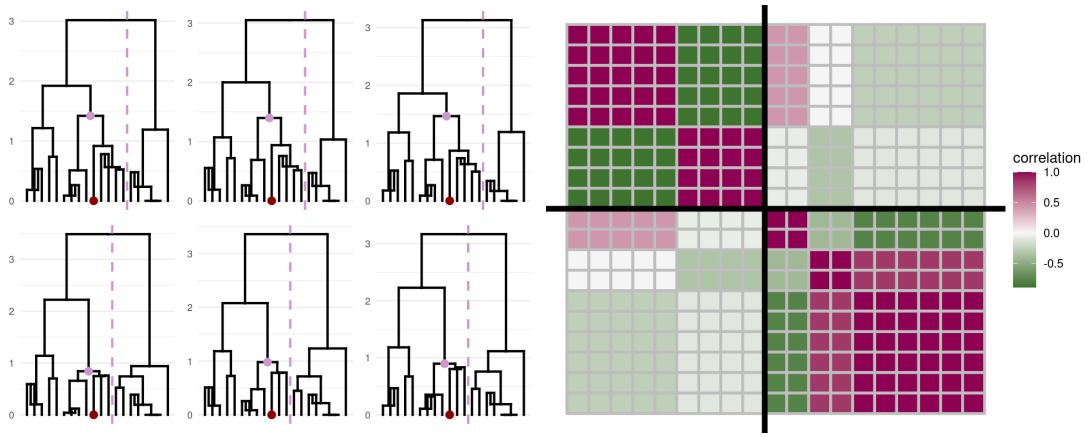

Fig. 5. Left: Dendrograms obtained for one simulation with parameters $n = 6$ and $B = 20$ (the simulation corresponding to the smallest p-value over the 1,000 simulations). The 3 dendrograms at the top correspond to the first condition and the 3 at the bottom to the second condition. Right: Subset of the empirical correlation matrix, $\widehat{\Sigma}$, for this simulation, based on cophenetic distance entries of leaf pairs of the form $\{(j^*, k)\}_{k=1,\ldots,B}$ for a fixed j^* (highlighted by a red dot on the dendrograms of the left panel and by a black line on the matrix of the right panel).

onal Hotelling test, the one based on matrices is equivalent (setting $n = 100$, $B = 20$) or much worse (other two settings) than/to the one based on cophenetic distances. This illustrates the robustness of the tree model for this type of hierarchical data.

As expected, the permutation test controls the type-I-error. It is the closest to uniformly distributed p-values under H_0 in the setting "small trees – moderate sample

size" ($n = 100$, $B = 20$). However, our experiments illustrate two major practical limitations of this test: its computational time and its conservativeness at low sample sizes. First, we were not able to perform this test for the "large trees – moderate sample size" setting ($n = 40$, $B = 100$) because of an excessive computational time (longer than 2 days). Indeed, this test requires the inversion of a matrix of dimension $p = B(B-1)/2$, and thus has a computation cost of the order $\mathcal{O}(B^6)$. This complexity is prohibitive even for moderate B, especially when several of these inversions have to be performed as in a permutation test. In addition to a prohibitive computational time even for a small number of permutations (see Figure 6 top left), this test also lacks resolution in the standard "small sample size" setting. Indeed, the maximum number of distinct permutations for a sample size of n is $P = \binom{n}{n_1}$ ($P = 20$ for $n_1 = n_2 = 3$). Therefore, the permutation p-values are multiples of $1/P$, as illustrated by the stepwise ECDF in Figure 4 (top left). In particular, the smallest p-value that can be achieved by a permutation test with P different permutations is $1/P$ (and even equal to $2/P = 0.1$ for $n_1 = n_2 = 3$ because, in the balanced case, some permutations are exactly symmetric). This lack of resolution implies that this test is overly conservative in this setting: regardless of the data, it is not possible to declare a test significant at a level less than 0.1. Note that this issue is even more problematic when several tests (on different sets of matrices) are performed (as in Section 4.3): in such a situation, it is virtually impossible for this test to detect signal due to the necessary multiple testing correction.

Finally, the results of the tree test and matrix test are very close. Both methods appropriately control the type-I error rate except for the most challenging simulation setting ($n = 6$ and $B = 20$) for which the number of p-values below 5% is, respectively, 9.9% and 10.8% for the tree test and matrix test. We note that, in these tests, the p-value ECDF is systematically closer to the diagonal for the tree test compared to the matrix test, suggesting that the matrix test is more conservative. However, no definitive conclusion can be drawn from these simulations only assessing the type-I error rate (this rate is smaller for the tree test in the $n = 100$, $B = 20$ simulation setting but larger in the $n = 40$, $B = 100$ simulation setting; values shown in the companion code repository, see information in Section "Data and code availability").

Figure 6 compares the computation times of the different methods‡, and shows that computation times for the tree test and matrix test are comparable. Interestingly, this indicates that the first step of the tree test (described in Section 2.1) comes at a negligible computational cost. Note that the number of permutations for the permutation test

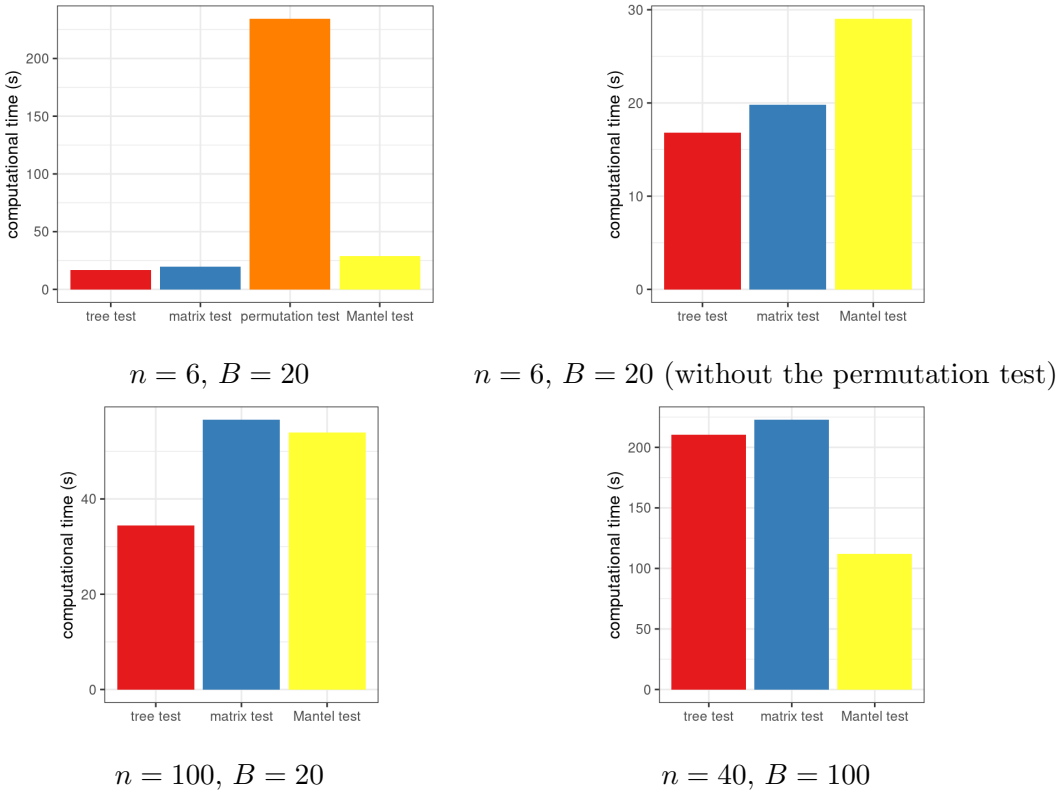

Fig. 6. GWAS: Comparison of the computation time for the different methods. Each plot corresponds to a choice of the parameters n (total number of trees in the two conditions) and B (number of leaves in the trees). Only one plot includes the permutation test (top right) that has a computation time very large compared to the other methods.

was set to 1,000, except for the case $n = 6$ where we only could use the 10 distinct

‡We did not represent the computation time for the permutation test in the $n = 100$, $B = 20$ setting because it was also much too large compared to the three other methods to be seen. Also, we did not represent the computation time of the diagonal Hotelling test, because the results of these tests can be directly derived from the computations performed for the tree test (so the computation times are the same for both tests).

permutations available. Even restricted to 10 permutations, it remained the method with the largest computation time ($\sim$ 8 times larger than the second one, see Figure 6 top left.).

Additional results (including p-value distribution and test statistics distribution for all tests) are provided in the notebooks of our companion code repository.

4.3. Differential analysis of Hi-C experiments

To show how the tree test can be applied in the field of 3D genomics, we used Hi-C data from a previous study (Marti-Marimon et al., 2021).

The Hi-C protocol aims to identify genomic regions that are located nearby in the 3D space within the cell nucleus, providing useful insights into fundamental biological functions (Lupiáñez et al., 2015; Won et al., 2016; Zheng and Xie, 2019). The generated information is a sparse square matrix whose entries correspond to the number of observed contacts between any given pair of genomic regions (or "bins"). From a mathematical point of view, Hi-C matrices can be interpreted as matrices of similarities between bins.

In this section, we used Hi-C matrices coming from 6 experiments performed on muscle tissues from 6 different pig fetuses. Each experiment produced 18 Hi-C matrices, one per pig chromosome (excluding X and Y). Here, the matrix resolution is 200 kb, which means that chromosomes are segmented into $200,000$ bases long bins.

The 6 pig fetuses originated from two biological conditions related to their maturity development at the end of gestation: 3 at 90 days of gestation and 3 at 110 days of gestation. Important modifications of the genome conformation have been observed between these two stages (Marti-Marimon et al., 2021), stressing the need for statistically relevant comparison methods. To date, standard methods for differential analysis of Hi-C data mostly proceed with tests at the bin pair level (Djekidel et al., 2018; Ardakany et al., 2019; Stansfield et al., 2019; Cook et al., 2020) and return positive results scattered in the similarity matrix, which are not easily interpretable biologically. In contrast, the tree test focuses on the hierarchical organization of genomic regions that can cover several contiguous bins, therefore capturing differences between large functional structures like TADs for instance (Dixon et al., 2012).

At the largest scale though, testing entire chromosomes for significant differences is of limited biological interest due to the lack of resolution. In order to identify genomic regions with significant structural modifications between the two stages, we applied the tree test on this dataset for 743 submatrices ("clusters") corresponding to a partition of the original chromosome matrices (the same partition is used across experiments; the way data were preprocessed and submatrices were chosen is described in Appendix B).

As a control for our test procedure, we performed the same test after permuting the condition labels identically for all the bin pairs of the same cluster (different label permutations were realized across clusters). Note that we performed one permutation per cluster to assess the p-value distribution under permutation. This is not a permutation test, where multiple permutations would be needed for each tested cluster. Furthermore, these 743 tests cannot be considered as independent, due to possible dependencies between neighboring genome regions. This is quite different from the preceding section where we performed independent replications of the same comparison in a controlled setting. The empirical distribution of raw p-values for the original and permuted data are shown in Figure 7 for all methods described in Section 4.1, except for the permutation test. Indeed, as explained above (see Figure 4; top left), the granularity of the p-values implied by the permutation framework makes the permutation test powerless for very low sample sizes, especially in a multiple testing context. The figure also gives, for all methods, the percentage of tests declared significant for adjusted p-values (Benjamini & Hochberg (BH) procedure; Benjamini and Hochberg (1995)) smaller than 5%.

In accordance with the results of Section 4.2, Hotelling tests yield a large inflation of false positives (from 21% to almost 100% of positive discoveries in the permuted setting). On the contrary, the tree test and the matrix test both properly control the type-I error rate. However, the number of positive detections in the non permuted case is strikingly different between these two tests and suggests that the power of the tree test is much larger than that of the matrix test. Indeed, the matrix test could not detect any significant difference whereas the tree test declares approximately one third of the genome (34.2%) as significantly different between stages. The latter conclusion is consistent with previous findings in genome conformation plasticity (Dixon et al., 2015),

and for this dataset in particular (Marti-Marimon et al., 2021), which suggests that the tree test is more appropriate than the matrix test.

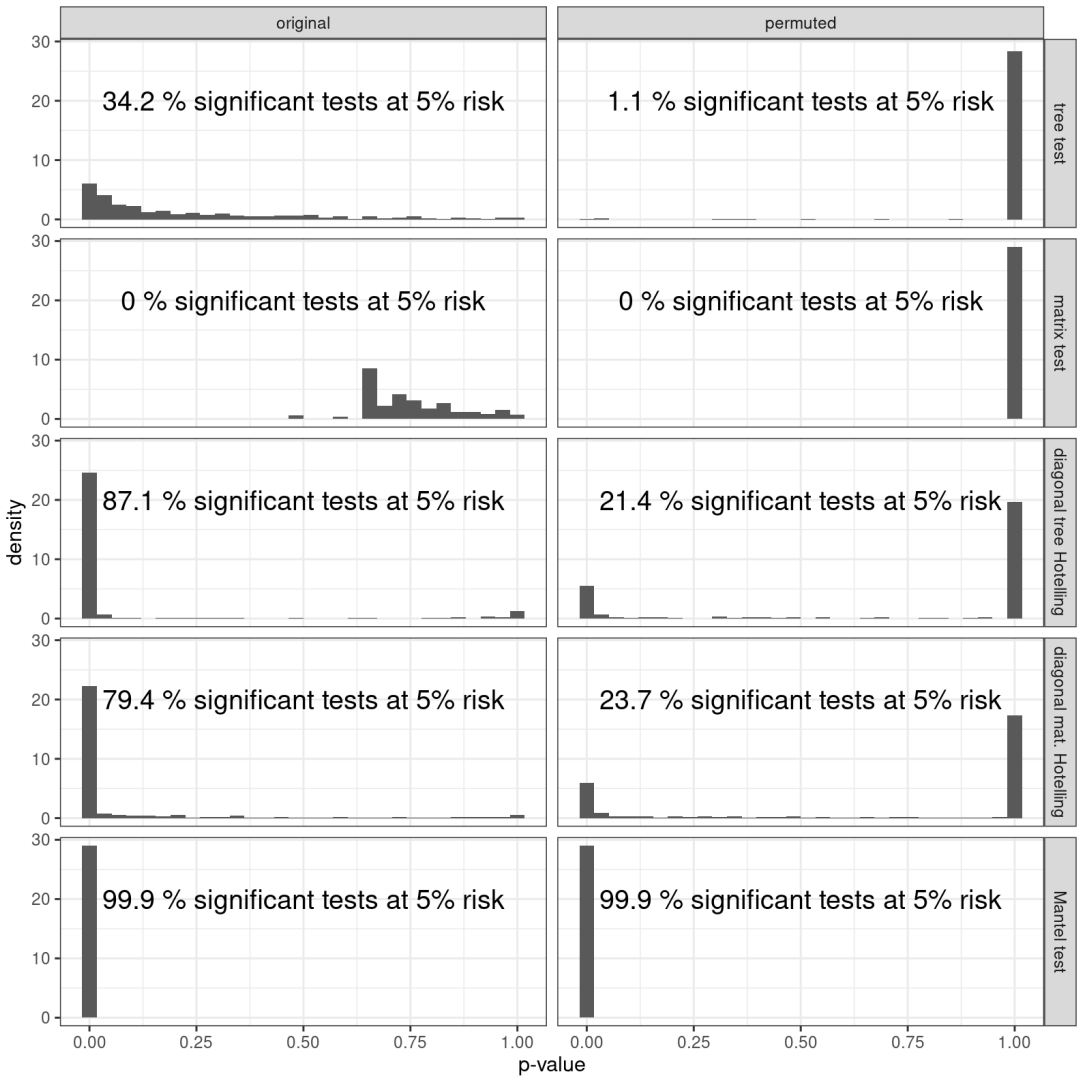

Fig. 7. p-value distribution for the different differential analysis procedures: original data (left) and permuted conditions (right). The percentage of tests declared significant is based on BH-adjusted p-values smaller than 5%.

To further confirm the validity of the the tree test, the empirical distribution of the individual statistics (as in Equation (2)) was compared with their theoretical distributions under H_{0j}. For the tree test, variance squeezing resulted in an increase of $\nu_0 = 1.17$ and 1.20 in the degrees of freedom for the original and permuted dataset, respectively. The null distribution is thus expected to be a Student with $\text{df} = 6 + \nu_0 = 7.17$ and 7.20, respectively. As the difference between these two distributions is negligible, empirical test statistic distributions for the original and permuted data were both compared to $\mathcal{T}_{7.17}$. The same was done for the matrix test and the comparison is displayed in Figure 8. As noted in the preceding paragraph, the underlying tests are most probably not independent.

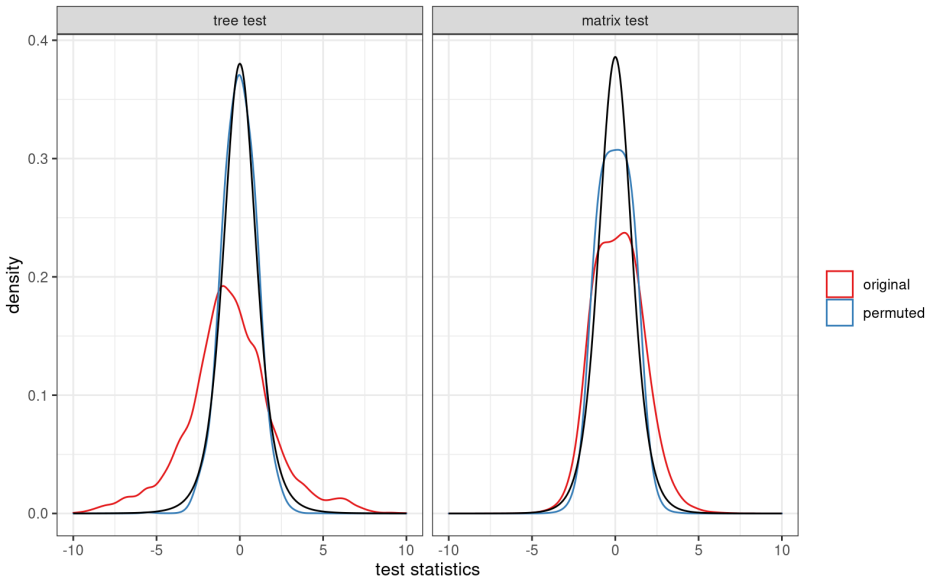

Fig. 8. Empirical distribution of bin pair statistics (as in Equation (2)), for the original data (red) and the permuted data (blue) compared to the theoretical null distribution (black). Results for the tree test are on the left and results for the matrix test are on the right.

This confirmed the consistency between the results of the tree test on the permuted dataset with the theory, and supported the validity of positive results. In contrast, the matrix test shows a more pronounced departure from the expected distribution under the permutation setting, suggesting a possible unsuitability of the test to this type of

data. On a side note, the observed statistics for the tree test statistics are shifted to the left when compared to the theoretical distribution (the left panel of Figure 8), which indicates that cophenetic distances are overall larger at 110 days of gestation than at 90 days of gestation. This suggests a higher degree of chromatin compaction at 90 days of gestation globally, which is again consistent with previous results from the same dataset (Marti-Marimon et al., 2021).

4.4. Comparison study for simulated biological and technical replicates

To further investigate the differences between the tree test and the matrix test, we used the same Hi-C data from Marti-Marimon et al. (2021) as before, but in a simulation setting. Artificial technical and biological replicates were generated to assess the benefit and robustness of the tree model. Various amounts of differences between replicates were obtained as follows:

- 2×3 **simulated technical replicates (TR)** were generated by independently removing 20% of the counts from the same original matrix (first replicate of the condition "110 days"; uniform subsampling). These replicates were randomly assigned to two simulated conditions. No signal is expected between these conditions because the generation process exactly corresponds to the inherent technical noise of the data;

- 2×3 **simulated technical replicates from any matrix (TRAll)** were generated using the same subsampling process on each of the original matrices separately. These replicates were randomly assigned to two simulated conditions. Again, no signal is expected between these conditions but this allows to maintain some of the original variability across replicates;

- 2×3 **simulated biological replicates from the same condition (BR1cond)** were generated using the same subsampling process on two matrices from the same condition (first and second replicates of the condition "110 days"). All matrices obtained from one of the two original replicates were assigned to one of the two simulated conditions. No strong signal is expected from this experiment since the

generation process corresponds to the biological difference between two different animals at the same developmental stage;

- 2×3 **simulated biological replicates from the two different conditions (BR2cond)** were generated using the same subsampling process on two matrices, one from each condition (first replicate of each condition). All matrices obtained from one of the two original matrices were assigned to one of the two simulated conditions. In this situation, a stronger signal is expected between conditions since the generation process starts from animals at two different developmental stages;

- 2×3 **simulated biological replicates from the two different conditions using all original matrices (BR2condAll)** were generated using the same subsampling process on each of the original matrices separately, maintaining the same condition assignment. In this situation, the same strong signal is expected between the two conditions but with a larger variability within conditions than in the previous simulation.

The simulation process is illustrated in Figure 9. Simulations were conducted similarly to Section 4.2 and 500 simulated experiments were generated on each of the 81 clusters from chromosome 1 (whose sizes ranged from 6 to 27 bins). Within each simulation type, run and method, adjusted p-values were computed using the BH procedure.

Figure 10 shows the p-value distribution for the tree test (top row) and the matrix test (bottom row), across the five scenarios considered (in rows). The percentage of "significant" tests is also indicated in each panel. For the H_0 scenarios (**TR** and **TRAll**, first two columns), we report the proportion of p-values below 5% with no multiple testing adjustment, in order to assess type-I error control. Since the other three scenarios involve comparisons between biological conditions, we expect differences between conditions to be detected in these cases. Therefore, we report the proportion of BH-adjusted p-values below 5%, in line with what would be done in real-life situation in order to control FDR. These results show that:

- for the technical replicate (**TR** and **TRAll**) simulations, both the tree test and the matrix test have a false positive rate below the expected value (5%). However,

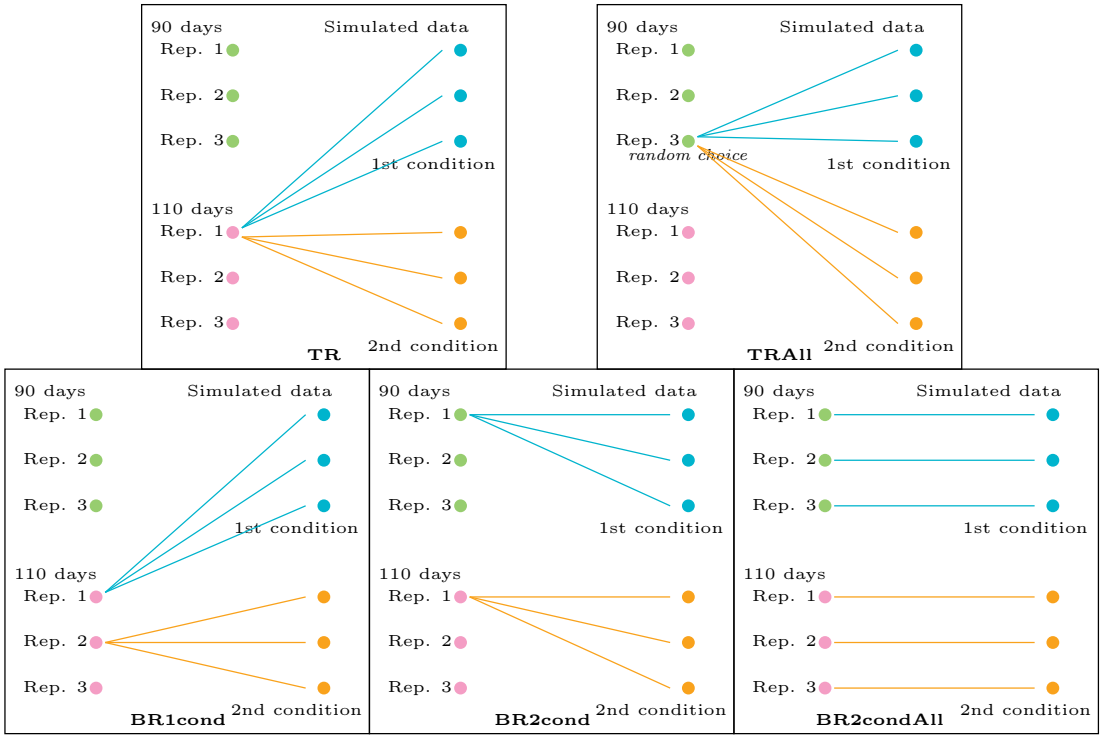

Fig. 9. Illustration of the simulation process from pig Hi-C data. For each setting, the left part of the figure represents the real data (with 3 replicates in two conditions) and the right part of the figure represents the simulated data (also with 3 replicates in two conditions). The straight lines between these two parts represent the subsampling process within matrices.

the matrix test is overly conservative, with no false discoveries at all in the **TRAll** setting for a thresholding of adjusted p-values at 5%. These results are consistent with those obtained in Section 4.2;

- for the biological replicate (**BR1cond** and **BR2cond**) simulations, both tests obtain a higher ratio of positive results, as expected. However, the matrix test behaves similarly in both cases, whereas **BR1cond** is expected to show weaker differences than **BR2cond**. At a more conservative threshold (0.1%), the difference between the two tests is even more pronounced: the tree test detects 55.9% and 67.4% of positive results for **BR1cond** and **BR2cond** respectively, whereas matrix test continues to detect 100% of positive results in both settings. This indicates a high

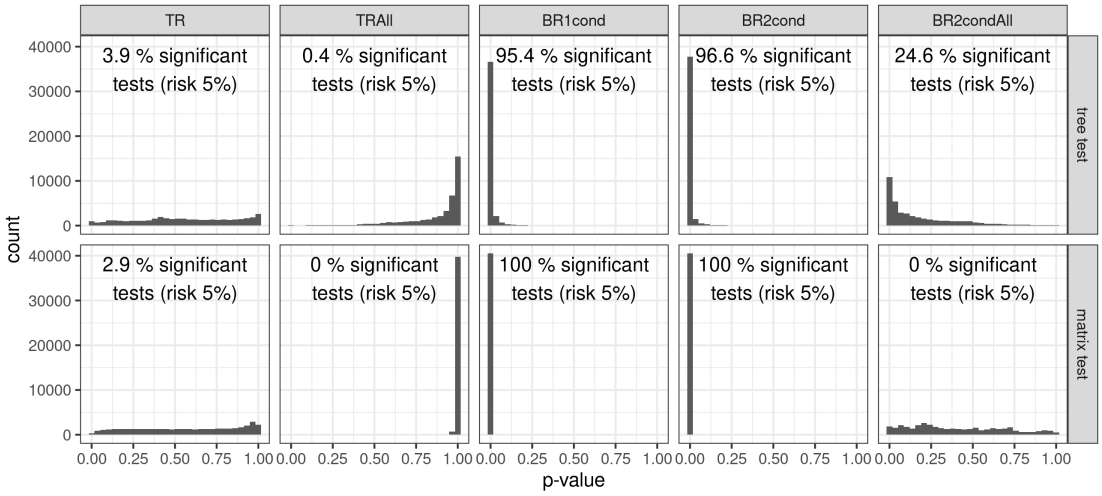

Fig. 10. Simulated data based on the pig HiC dataset. Distribution of adjusted p-values and percentage of tests detected as positive by controlling either p-values (H_0 settings **TR** and **TRAll**) or adjusted p-values (other settings where some differences between the two conditions are expected) at a level of 5%.

sensitivity of matrix test to differences between conditions and a possible inflation of false positive detections when groups are particularly homogeneous;

- more importantly, in the **BR2condAll** setting, which is the closest to the true situation, the matrix test detects no positive result whereas the tree test declares about one fourth of tests as positive. This result is highly consistent with the results obtained in Section 4.3 and suggests that, unlike the tree test, the matrix test does not cope well with biological variability within a condition.

5. Conclusion

We have introduced a tree-based method to test for significant differences between two sets of similarity matrices obtained from two different conditions. The data representation obtained at the first step of this method explicitly incorporates the hierarchical structure present in the input matrices. Then, our procedure is the result of an aggregation of univariate moderated tests, designed to cope with the typical high-dimensional

setting encountered in genomic studies. Unlike the multivariate approaches described in Section 3.2, the tree test does not need to estimate the inverse of a covariance matrix nor does it assume independence between tests. Moreover, it does not require intensive computation since it is based on univariate parametric tests.

Methodological choices imply tradeoffs, and we acknowledge that our selected choices may not always be optimal in all contexts, nor can we guarantee that the theoretical assumptions ensuring the formal validity of our test, such as the Gaussianity of the marginal distribution of $\mathbf{X}_j$, and PRDS property for the test statistics, always hold. In particular, it is not obvious a priori that our proposed approach is superior to non-parametric alternatives (*e.g.*, permutation tests), to multivariate tests, or to tests that do not take the hierarchical structure of the input data into account. Therefore, the extensive numerical experiments performed in Section 4 are an important contribution of this work. These experiments illustrate the relevance of our choices as compared to these alternatives, both in terms of type-I error control and in terms of statistical power, in particular in the context of very low sample sizes, which are typical of genomic studies.

In practical situations such as the GWAS and Hi-C data analysis contexts that motivated this work, users might want to detect the largest or the smallest subset of leaves leading to differential structures between the two conditions rather than only performing a test at the global tree level. In the Hi-C application described in Section 4.3, this problem is addressed by performing tests on subtrees corresponding to pre-determined clusters of the chromosome wide trees. Automatically detecting regions of interest to test from the data could facilitate the adoption of our proposed test by the genomic community. Such an extension raises challenging statistical and computational issues, which constitute an interesting perspective for future works.

Acknowledgments

The authors gratefully acknowledge the INRAE/Inria doctoral program 2018 for the funding of N.R.'s PhD thesis of N.R. and the CNRS (Mission "Osez l'interdisciplinarité") for funding the SCALES project. The authors are grateful to the "Genotoul-bioinfo" platform (INRAE Toulouse, `http://bioinfo.genotoul.fr/`) and its staff for providing

computing facilities.

Conflict of Interest

The authors have no conflict of interest to declare.

Data and code availability

All scripts for the described simulations are accessible from the GitLab repository `https://forgemia.inra.fr/scales/differential-analysis-of-trees`. Scripts have been run on the genologin cluster facility provided by the Bioinformatics platform "Genotoul-bioinfo" `http://bioinfo.genotoul.fr`. All scripts have been run using R version 4.0.2 (R Core Team, 2020) with the environment setup as in the **renv** (Ushey, 2020) file provided in the repository. Explicitly loaded packages are: **adjclust** (Ambroise et al., 2019), **ape** (Paradis and Schliep, 2019), **arrangements**, Bioconductor/**csaw** (Lun and Smyth, 2016), **data.table**, **dendextend** (Galili, 2015), **future** and **future.apply**, **ggplot2** (Wickham, 2016), **ggpubr**, **gridExtra**, **Hotelling**, Bioconductor/**limma** (Ritchie et al., 2015), **RColorBrewer**, **reshape** (Wickham, 2007), Bioconductor/**snpStats**, **tidyverse** (Wickham et al., 2019).

Hi-C count matrices used in Sections 4.3 and 4.4 are available on INRAE data portal with DOI `https://doi.org/10.15454/8BLMNQ`. These count matrices have been obtained from raw sequencing data available on ENA portal `https://www.ebi.ac.uk/ena/` under accession number PRJEB40576 as documented in Marti-Marimon et al. (2021) and in Appendix B.

Appendix

A. Moderated variance for Student's tests

As initially proposed by Smyth (2004), we follow an empirical Bayes framework in which σ_j^{-2} is assumed to be a random variable with a scaled inverse χ^2 distribution with parameters ν_0 and σ_0^2: $\nu_0\sigma_0^2/\sigma_j^2 \sim \chi^2_{\nu_0}$, where χ^2_ν denotes a χ^2 distribution with ν degrees of freedom.

The posterior distribution of σ_j^2 can be shown to also follow a scaled inverse χ^2 with parameters $\nu_0 + n$ and $(\nu_0\sigma_0^2 + (n-2)\tilde{\sigma}_j^2)/(\nu_0 + n)$. In particular, we have

$$\mathbb{E}(\sigma_j^2|\tilde{\sigma}_j^2) = \frac{\nu_0\sigma_0^2 + (n-2)\tilde{\sigma}_j^2}{\nu_0 + n - 2}. \tag{7}$$

This result is already known (see e.g., Smyth (2004)) but we were not able to find a proof for it in the literature so we give it below for completeness.

PROOF. Let $\nu := \nu_j = n - 2$ be the number of degrees of freedom of the frequentist Student's statistic. We also introduce additional notations from Smyth (2004): For $j = 1, \ldots, p$, let $\beta_j := \mu_{1j} - \mu_{2j}$, and $\hat{\beta}_j := \overline{\mathbf{X}}_j^{(1)} - \overline{\mathbf{X}}_j^{(2)}$ its natural estimator.

LEMMA 1. *If, for $k \in \{1, 2\}$, $(\mathbf{X}_{ij})_{i \in \mathcal{C}_k}$ are i.i.d. observations of $\mathcal{N}(\mu_{kj}, \sigma_j^2)$,*

(a) the distribution of $\hat{\beta}_j$ given β_j and σ_j^2 is given by:

$$\hat{\beta}_j|\beta_j, \sigma_j^2 \sim \mathcal{N}\left(\beta_j, \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\sigma_j^2\right),$$

(b) the distribution of $\tilde{\sigma}_j^2$ given σ_j^2 is given by

$$\tilde{\sigma}_j^2|\sigma_j^2 \sim \frac{\sigma_j^2}{\nu}\chi_\nu^2.$$

See e.g., Saporta (1990) for a proof of Lemma 1.

In the Bayesian framework defined by Smyth (2004), Lemma 1 implies that:

- the likelihood of the model is given by $p(\tilde{\sigma}_j^2|\sigma_j^2)$. As the probability distribution of $X = aZ$ for $a > 0$ and $Z \sim \chi_\nu^2$ is given by

  $$p_X(x) = \frac{(1/2)^{\nu/2}}{\Gamma(\nu/2)}\, a^{-\nu/2}\, x^{\nu/2-1} e^{x/(2a)},$$

  we thus have that

  $$p(\tilde{\sigma}_j^2 = x|\sigma_j^2 = y) = K y^{-\nu/2} x^{\nu/2-1} e^{-(\nu x)/(2y)}$$

  where K is a normalization constant not depending on x or y,

- the prior distribution of σ_j^2 satisfies $\frac{1}{\sigma_j^2} \sim \frac{1}{\nu_0\sigma_0^2}\chi_{\nu_0}^2$, that is, $\sigma_j^2 \sim \nu_0\sigma_0^2\,\mathrm{Inv}\chi_{\nu_0}^2$. As the probability distribution of $Y = bZ'$ for $b > 0$ and $Y \sim \mathrm{Inv}\chi_{\nu_0}^2$ is given by

  $$\frac{2^{-\nu_0/2}}{\Gamma(\nu_0/2)}\, b^{\nu_0/2}\, y^{-\nu_0/2-1} e^{-b/(2y)},$$

we have that

$$p(\sigma_j^2 = y) = K' y^{-\nu_0/2-1} e^{-(\nu_0\sigma_0^2)/(2y)}.$$

Omitting terms not depending on y, the posterior probability distribution is proportional to

$$\begin{aligned} p(\sigma_j^2 = y|\tilde{\sigma}_j^2) &\sim \text{Likelihood} \times \text{Prior} \\ &= y^{-\nu/2}\, e^{-(\nu\tilde{\sigma}_j^2)/(2y)}\, y^{-\nu_0/2}\, e^{-(\nu_0\sigma_0^2)/(2y)} \\ &= y^{-(\nu+\nu_0)/2}\, e^{-(\nu\tilde{\sigma}_j^2+\nu_0\sigma_0^2)/(2y)} \\ &= y^{-(\nu+\nu_0+2)/2-1}\, e^{-(\nu\tilde{\sigma}_j^2+\nu_0\sigma_0^2)/(2y)} \end{aligned}$$

where we recognize the distribution of $(\nu\tilde{\sigma}_j^2 + \nu_0\sigma_0^2)\text{Inv}\chi^2_{\nu+\nu_0+2}$.

Equation (7) is a consequence of this result, as the expectation of $\text{Inv}\chi^2_{\nu+\nu_0+2}$ is $(\nu + \nu_0 + 2 - 2)^{-1} = (\nu + \nu_0)^{-1}$.

Then, following Smyth (2004), we consider the moderated test statistic t_j defined in Equation (2) associated to the "squeezed" variance estimate $\hat{\sigma}_j^2 = \mathbb{E}(\sigma_j^2|\tilde{\sigma}_j^2)$ given in (7). The distribution of this statistic under H_{0j} is $t_j|\mu_2 = \mu_1 \sim \mathcal{T}_{\nu_0+n-2}$, where $\mathcal{T}_\nu$ is the Student distribution with ν degrees of freedom. Thus, an "individual p-value" π_j for each H_{0j}, $j = 1, \ldots, p$ is obtained as:

$$\pi_j = 2(1 - F_{\nu_0+n-2}(|t_j|)),$$

where F_ν is the cumulative distribution function of $\mathcal{T}_\nu$.

In practice, ν_0 and σ_0 are obtained using the first two moments of $\log(\tilde{\sigma}_j^2)$ and α_0 using an average of the top order statistics of the t_j, as initially proposed by Smyth (2004) and implemented in the R package **limma** (Ritchie et al., 2015).

B. Preprocessing of the Hi-C matrices

Hi-C matrices used in Section 4.3 were processed by chromosome (the same preprocessing was performed for the 18 chromosomes independently). First, the six matrices were corrected for differences in sequencing depths across experiments using MA normalization (Cleveland and Devlin, 1988; Ballman et al., 2004) resulting in six corrected matrices for each chromosome.

Then, the corrected matrices were divided into submatrices of interpretable sizes using the procedure illustrated in Figure 11. The six matrices were first summed to obtain a merged matrix (1st step) on which constrained hierarchical clustering (HC) was performed (as implemented in **adjClust**; 2nd step). A relevant clustering was thus obtained by a cut of the resulting dendrogram at the number of clusters obtained using the broken stick heuristic (Bennett (1996); 3rd step).

The resulting submatrices (the red triangles in the six matrices at the bottom left side of Figure 11) were then submitted to different tests as described in Section 4.3. In tests using the moderated variance procedure, the variance was moderated at the genome-wide level (*e.g.*, using the variances computed for all clusters of all chromosomes) and the cluster p-values were also corrected at the genome-wide level.

References

Ambroise, C., Dehman, A., Neuvial, P., Rigaill, G. and Vialaneix, N. (2019) Adjacency-constrained hierarchical clustering of a band similarity matrix with application to genomics. *Algorithms for Molecular Biology*, **14**, 22.

Ardakany, A. R., Ay, F. and Lonardi, S. (2019) Selfish: discovery of differential chromatin interactions via a self-similarity measure. *Bioinformatics*, **35**, i145–i153.

Bai, Z. and Saranadasa, H. (1996) Effect of high dimension: by an example of a two sample problem. *Statistica Sinica*, **6**, 311–329.

Ballman, K., Grill, D., Oberg, A. and Therneau, T. (2004) Faster cyclic loess: normalizing RNA arrays via linear models. *Bioinformatics*, **20**, 2778–2786.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, **57**, 289–300.

Benjamini, Y. and Yekutieli, D. (2001) The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, **29**, 1165–1188.

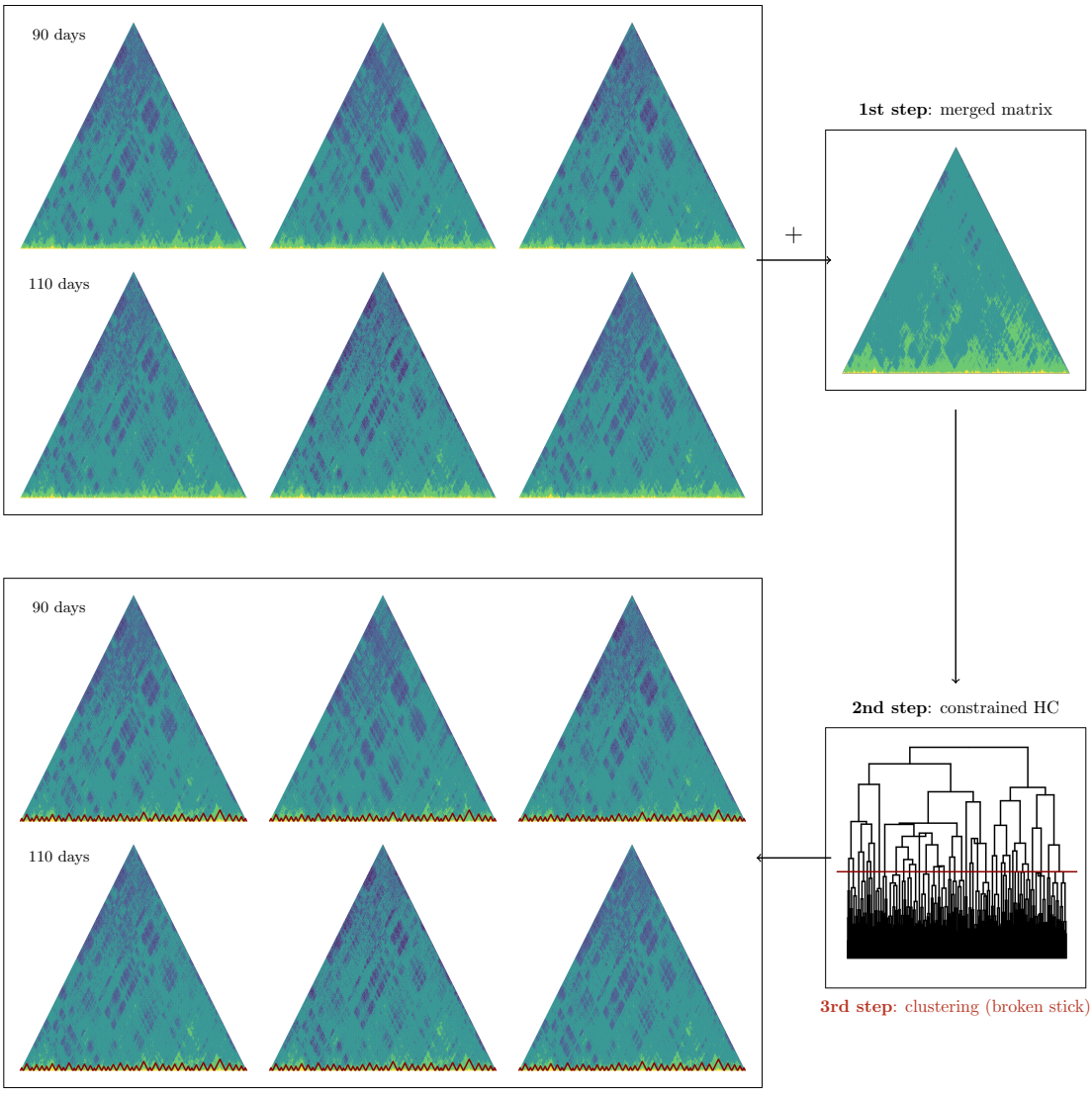

Fig. 11. Illustration of the sub-matrix definition for the Hi-C experiment conducted in Section 4.3. For each chromosome, the MA normalized matrices were first summed to obtained a consensus matrix. Using constrained hierarchical clustering combined with the broken stick heuristic, this merged matrix allowed to define common contiguous clusters, used to obtain sub-matrices from the six initial Hi-C matrices. This figure has been obtained using chromosome 7 of the dataset.

Bennett, K. D. (1996) Determination of the number of zones in a biostratigraphical sequence. *New Phytologist*, **132**, 155–170.

Billera, L. J., Holmes, S. P. and Vogtmann, K. (2001) Geometry of the space of phylogenetic trees. *Advances in Applied Mathematics*, **27**, 733–767.

Bordewich, M. and Semple, C. (2005) On the computational complexity of the rooted subtree prune and regraft distance. *Annals of Combinatorics*, **8**, 409–423.

Chakerian, J. and Holmes, S. (2012) Computational tools for evaluating phylogenetic and hierarchical clustering trees. *Journal of computational and Graphical Statistics*, **21**, 581–599.

Chen, L. S., Paul, D., Prentice, R. L. and Wang, P. (2011) A regularized Hotelling's t^2 test for pathway analysis in proteomic studies. *Journal of the American Statistical Association*, **106**, 1345–1360.

Chen, S. X. and Qin, Y.-L. (2010) A two-sample test for high-dimensional data with applications to gene-set testing. *Annals of Statistics*, **38**, 808–835.

Cleveland, W. and Devlin, S. (1988) Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association*, **83**, 596–610.

Cook, K. B., Hristov, B. H., Le Roch, K. G., Vert, J.-P. and Noble, W. S. (2020) Measuring significant changes in chromatin conformation with ACCOST. *Nucleic Acids Research*, **48**, 2303–2311.

DasGupta, B., He, X., Jiang, T., Li, M., Tromp, J. and Zhang, L. (1997) On distances between phylogenetic trees. In *Proceedings of the 8th annual ACM-SIAM Symposium on Discrete Algorithms (SODA '97)* (ed. M. Saks), 427–436. New Orleans, LA, USA: SIAM, Philadelphia, PA, USA.

Dixon, J. R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J. E., Lee, A. Y., Ye, Z., Kim, A., Rjagopal, N., Xie, W., Diao, Y., Liang, J., Zhao, H., Lobanenkov, V. V.,

Ecker, J. R., Thomson, J. A. and Ren, B. (2015) Chromatin architecture reorganization during stem cell differentiation. *Nature*, **518**, 331–336.

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S. and Ren, B. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**, 376–380.

Djekidel, M. N., Chen, Y. and Zhang, M. Q. (2018) FIND: difFerential chromatin INteractions Detection using a spatial Poisson process. *Genome Research*, **28**, 412–422.

Dong, K., Pang, H., Tong, T. and Genton, M. G. (2016) Shrinkage-based diagonal Hotelling's tests for high-dimensional small sample size data. *Journal of Multivariate Analysis*, **143**, 127–142.

Efron, B., Halloran, E. and Holmes, S. (1996) Bootstrap confidence levels for phylogenetic trees. *Proceedings of the National Academy of Sciences of the United States of America*, **93**, 13429–13434.

Fraser, J., Ferrai, C., Chiariello, A. M., Schueler, M., Rito, T., Laudanno, G., Barbieri, M., Moore, B. L., Kraemer, D. C., Aitken, S., Xie, S. Q., Morris, K. J., Itoh, M., Kawaji, H., Jaeger, I., Hayashizaki, Y., Carninci, P., Forrest, A. R., The FANTOM Consortium, Semple, C. A., Dostie, J., Pombo, A. and Nicodemi, M. (2015) Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. *Molecular Systems Biology*, **11**, 852.

Galili, T. (2015) dendextend: an R package for visualizing, adjusting, and comparing trees of hierarchical clustering. *Bioinformatics*, **31**, 3718–3720.

Goeman, J. J. and Solari, A. (2014) Multiple hypothesis testing in genomics. *Statistics in medicine*, **33**, 1946–1978.

Hébert, F., Causeur, D. and Emily, M. (2021) An adaptive decorrelation procedure for signal detection. *Computational Statistics & Data Analysis*, **153**, 107082.

Holmes, S. (2003a) Boostrapping phylogenetic trees: theory and methods. *Statistical Science*, **18**, 241–255.

— (2003b) Statistics for phylogenetic trees. *Theoretical Population Biology*, **63**, 17–32.

Hotelling, H. (1936) Relations between two sets of variates. *Biometrika*, **28**, 321–377.

Lun, A. T. and Smyth, G. K. (2014) De novo detection of differentially bound regions for ChIP-seq data using peaks and windows: controlling error rates correctly. *Nucleic Acids Research*, **42**, e95.

— (2016) csaw: a Bioconductor package for differential binding analysis of ChIP-seq data using sliding windows. *Nucleic Acids Research*, **44**, e45.

Lupiáñez, D., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J. M., Laxova, R., Santos-Simarro, F., Gilbert-Dussardier, B., Wittler, L., Borschiwer, M., Haas, S. A., Osterwalder, M., Franke, M., Timmermann, B., Hecht, J., Spielmann, M., Visel, A. and Mundlos, S. (2015) Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, **161**, 1012–1025.

Mallows, C. (1957) Non-null ranking models. I. *Biometrika*, **44**, 114–130.

Marti-Marimon, M., Vialaneix, N., Lahbib-Mansais, Y., Zytnicki, M., Camut, S., Robelin, D., Yerle-Bouissou, M. and Foissac, S. (2021) Major reorganization of chromosome conformation during muscle development in pig. *Frontiers in Genetics*, **12**, 748239.

Owen, M. and Provan, J. S. (2011) A fast algorithm for computing geodesic distances in tree space. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **8**, 2–13.

Paradis, E. and Schliep, K. (2019) ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, **35**, 526–528.

R Core Team (2020) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: `https://www.R-project.org/`.

Randriamihamison, N., Vialaneix, N. and Neuvial, P. (2021) Applicability and interpretability of Ward's hierarchical agglomerative clustering with or without contiguity constraints. *Journal of Classification*, **38**, 363–389.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W. and Smyth, G. K. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, **43**, e47.

Robinson, D. and Foulds, L. (1981) Comparison of phylogenetic trees. *Mathematical Biosciences*, **53**, 131–147.

Rødland, E. A. (2006) Simes' procedure is 'valid on average'. *Biometrika*, **93**, 742–746.

Saporta, G. (1990) *Probabilités, Analyses des Données et Statistique*. Editions Technip.

Shen, Y., Lin, Z. and Zhu, J. (2011) Shrinkage-based regularization tests for high-dimensional data with application to gene set analysis. *Computational Statistics & Data Analysis*, **55**, 2221–2233.

Simes, R. J. (1986) An improved bonferroni procedure for multiple tests of significance. *Biometrika*, **73**, 751–754.

Smyth, G. K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Methods in Genetics and Molecular Biology*, **3**.

Soler-Vila, P., Cuscó, P., Farabella, I., Di Stefano, M. and Marti-Renon, M. A. (2020) Hierarchical chromatin organization detected by TADpole. *Nucleic Acids Research*, **45**, e39.

Stansfield, J. C., Cresswell, K. G. and Dozmorov, M. G. (2019) multiHiCcompare: joint normalization and comparative analysis of complex Hi-C experiments. *Bioinformatics*. Forthcoming.

Steel, M. A. and Penny, D. (1993) Distributions of tree comparison metrics–some new results. *Systematic Biology*, **42**, 126–141.

The International HapMap Consortium (2003) The international HapMap project. *Nature*, **426**, 789–796.

Tong, T. and Wang, Y. (2012) Optimal shrinkage estimation of variances with applications to microarray data analysis. *Journal of the American Statistical Association*, **102**, 113–122.

Tusher, V. G., Tibshirani, R. and Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, **98**, 5116–5121.

Ushey, K. (2020) *renv: Project Environments*. URL: `https://CRAN.R-project.org/package=renv`. R package version 0.12.3.

Weinreb, C. and Raphael, B. J. (2016) Identification of hierarchical chromatin domains. *Bioinformatics*, **32**, 1601–1609.

Wickham, H. (2007) Reshaping data with the reshape package. *Journal of Statistical Software*, **21**.

— (2016) *ggplot2: Elegant Graphics for Data Analysis*. New York, USA: Springer-Verlag. URL: `https://ggplot2.tidyverse.org`.

Wickham, H., Averick, M., Bryan, J., Chang, W., D'Agostino McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Lin Pedersen, T., Miller, E., Milton Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K. and Yutani, H. (2019) Welcome to the tidyverse. *Journal of Open Source Software*, **4**, 1686.

Won, H., de La Torre-Ubieta, L., Stein, J. L., Parikshak, N. N., Huang, J., Opland, C. K., Gandal, M. J., Sutton, G. J., Hormozdiari, F., Lu, D., Lee, C., Eskin, E., Voineagu, I., Ernst, J. and Geschwind, D. H. (2016) Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature*, **538**, 523–527.

Won, S., Park, J.-E., Son, J.-H., Lee, S.-H., Park, B. H., Park, M., Park, W.-C., Chai, H.-H., Kim, H., Lee, J. and Lim, D. (2020) Genomic prediction accuracy using haplotypes

defined by size and hierarchical clustering based on linkage disequilibrium. *Frontiers in Genetics*, **11**.

Zheng, H. and Xie, W. (2019) The role of 3d genome organization in development and cell differentiation. *Nature Reviews Molecular Cell Biology*, **20**, 535–550.