IWTomics: testing high-resolution sequence-based "Omics" data at multiple locations and scales

Supplementary material

Marzia A. Cremona[1,†], Alessia Pini[2,†], Fabio Cumbo[3,4], Kateryna D. Makova[5,6], Francesca Chiaromonte[1,5,7,*] and Simone Vantini[2,*]

[1]Department of Statistics, The Pennsylvania State University, University Park, PA, USA
[2]MOX - Modeling and Scientific Computing, Department of Mathematics, Politecnico di Milano, Milano, Italy
[3]Department of Engineering, Third University of Rome, Rome, Italy
[4]Institute for Systems Analysis and Computer Science "Antonio Ruberti", National Research Council of Italy, Rome, Italy
[5]Center for Medical Genomics, The Huck Institutes of the Life Sciences, The Pennsylvania State University, University Park, PA, USA
[6]Department of Biology, The Pennsylvania State University, University Park, PA, USA
[7]Sant'Anna School of Advanced Studies, Pisa, Italy

[†]These authors contributed equally to this work
[*]To whom correspondence should be addressed: fxc11@psu.edu, simone.vantini@polimi.it

1 Interval-Wise Testing (IWT)

Let $y_{1,i}(x)$, $i = 1, \dots, n_1$ be the feature curves corresponding to the n_1 genomic regions of the first group, and $y_{2,j}(x)$, $j = 1, \dots, n_2$ be the feature curves corresponding to the n_2 genomic regions of the second group. We assume that $y_{1,i}(x)$ and $y_{2,i}(x)$ are two random samples from two independent random functions, defined in the interval I.

The IWT tests the null hypothesis H_0^I that the distributions of two random functions on the interval I are equal, versus the alternative hypothesis H_1^I that they differ. Let S be a subinterval $S = (x_a, x_b) \subseteq I$ or a complementary subinterval $S = I \setminus (x_a, x_b)$. We indicate with H_0^S and H_1^S the restrictions of the null and alternative hypotheses to S.

The first step of the IWT consists of a functional permutation test for H_0^S vs H_1^S on every possible $S = (x_a, x_b) \subseteq I$ and $S = I \setminus (x_a, x_b)$. Given a test statistic T, we estimate its permutational distribution under H_0^S by evaluating its value $T(S)$ for all possible permutations of the $n_1 + n_2$ curves. The test p-value p^S is the proportion of permutations that lead to a test statistic greater than or equal to the one observed in the actual data. When the number of possible permutations is large, we approximate the test p-value considering a fixed number B of random permutations.

The second step of the IWT concerns the computation of the adjusted p-value curve $\tilde{p}(x)$, defined at each $x \in I$ as

$$\tilde{p}(x) = \sup_{S \ni x} p^S$$

The adjusted p-value curve $\tilde{p}(x)$ controls the interval-wise error rate, i.e. the probability of rejecting the null hypothesis H_0^S on every interval $S \subseteq I$ where it is true:

$$\forall S \subseteq I \text{ s.t. } H_0^S \text{ is true} \qquad \Pr\left[\forall x \in S, \tilde{p}(x) \leq \alpha\right] \leq \alpha.$$

The final step of the IWT consists of identifying locations with a significant difference between the two groups by selecting all the points $x \in I$ such that $\tilde{p}(x) < \alpha$, where α is the desired significance level.

1.1 Testing multiple locations and scales

In order to detect both locations and scales at which differences between the random functions are significant, the extended version of IWT evaluates multiple scales and computes an adjusted p-value curve $\tilde{p}_s(x)$ for each scale $s \leq |I|$ (the size of I). For each fixed scale s, $\tilde{p}_s(x)$ is computed considering only subintervals $S = (x_a, x_b) \subseteq I$ and complementary subintervals $S = I \setminus (x_a, x_b)$ of length $|S| \leq s$. As a consequence, $\tilde{p}_s(x)$ controls the interval-wise error rate on all intervals of length at most s and identifies locations that are significant at scale s.

In practice, feature curves $y_{1,i}(x)$ and $y_{2,i}(x)$ are observed only on a discrete grid of equally spaced points $x_1, \ldots, x_K$ (each point corresponds to a measurement of the feature in one small window). As a consequence, the algorithm implemented by *IWTomics* partitions the interval I in K subintervals $S_1, \ldots, S_K$ of the same size – each containing exactly one point of the grid. The test is performed considering these K subintervals as locations, and the possible scales ranges from 1 (1 subinterval, no adjustment applied) to K (K subintervals, adjustment applied up to the entire interval I). The adjustment performed at scale s takes into considerations the hypotheses on every single subinterval $H_0^{S_1}, \ldots, H_0^{S_K}$, the hypotheses on every interval obtained as the union of 2 subintervals $H_0^{S_1 \cup S_2}, \ldots, H_0^{S_{K-1} \cup S_K}$, and so on up to every interval obtained as union of s subintervals $H_0^{S_1 \cup \cdots \cup S_s}, \cdots, H_0^{S_{K-s+1} \cup \cdots \cup S_K}$. The adjustment performed by *IWtomics* obviously induces a loss of statistical power with respect to a global test on the entire interval I, but it allows one to identify locations and scales at which the two random functions differ significantly. In addition, this adjustment retains better statistical power than procedures such as Bonferroni or Benjamini-Hochberg applied to the points of the grid. A visual comparison between the unadjusted p-value curve $\tilde{p}_1(x)$ (at scale 1) and the adjusted p-value curve $\tilde{p}_s(x)$ for a given scale, provides the user with an indication about the loss of statistical power due to adjustment.

1.2 Test statistics

In the extended version of the IWT, different test statistics can be employed in the first step to compute the test p-value p^S, allowing one to focus on different characteristics of the curve distributions.

For every $S \subseteq I$ we define the *mean* test statistic as

$$T_{mean}(S) = \frac{1}{|S|} \int_S \Big(\bar{y}_1(x) - \bar{y}_2(x)\Big)^2 \mathrm{d}x,$$

where $\bar{y}_1(x) = \frac{1}{n_1} \sum_{i=1}^{n_1} y_{1,i}(x)$ and $\bar{y}_2(x) = \frac{1}{n_2} \sum_{j=1}^{n_2} y_{2,j}(x)$ are the sample means of the curves in the two groups.

The *median* test statistic is defined as

$$T_{median}(S) = \frac{1}{|S|} \int_S \Big(y_1^{(0.50)}(x) - y_2^{(0.50)}(x)\Big)^2 \mathrm{d}x,$$

where $y_1^{(0.50)}(x)$ and $y_2^{(0.50)}(x)$ are the pointwise sample medians of the curves in the two groups.

Similarly, the *multi-quantile* test statistic is given by

$$T_{multi-quantile}(S) = \sum_{q \in Q} \frac{1}{|S|} \int_S \Big(y_1^{(q)}(x) - y_2^{(q)}(x)\Big)^2 \mathrm{d}x,$$

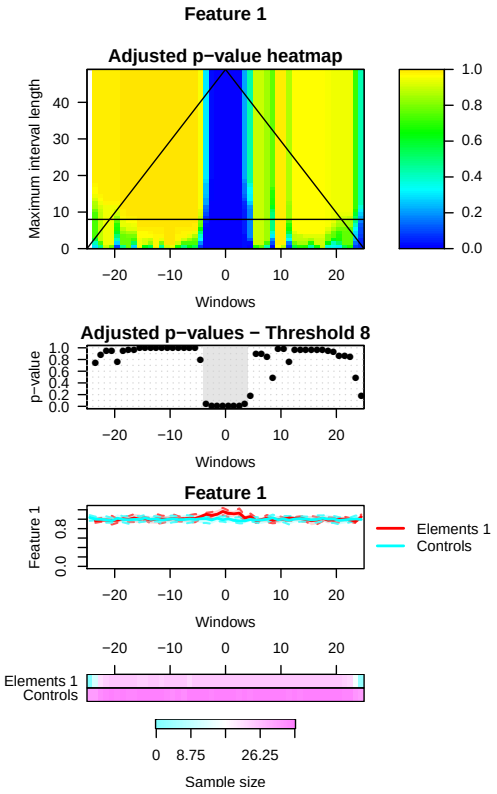
where, for every $x \in I$, $y_1^{(q)}(x)$ and $y_2^{(q)}(x)$ are the quantiles of order q of the curves in the two groups, and Q is a set of probabilities.

Finally, we define the *variance* statistic as

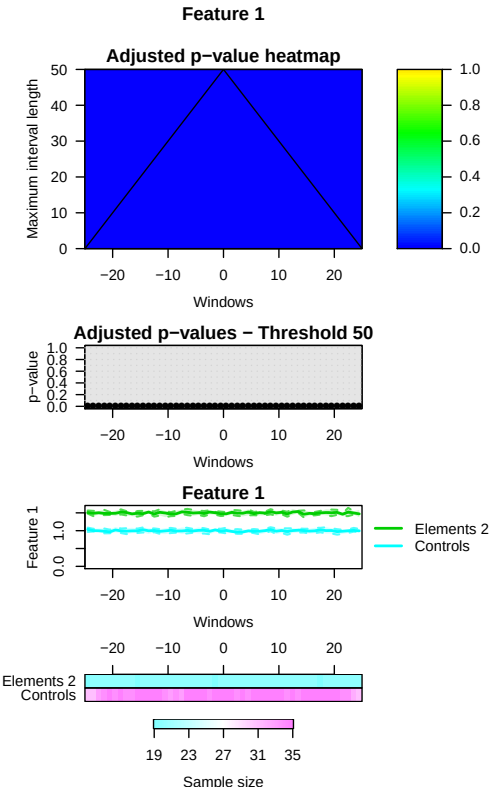
$$T_{variance}(S) = \frac{1}{|S|} \int_S \frac{\mathrm{Var}(y_1(x))}{\mathrm{Var}(y_2(x))} \mathrm{d}x,$$

where $\mathrm{Var}(y_1(x))$ and $\mathrm{Var}(y_2(x))$ are the pointwise sample variances of the curves in the two groups. In this case the test p-value p^S is computed considering both tails of the permutational distribution, in order to obtain a two-sided test.

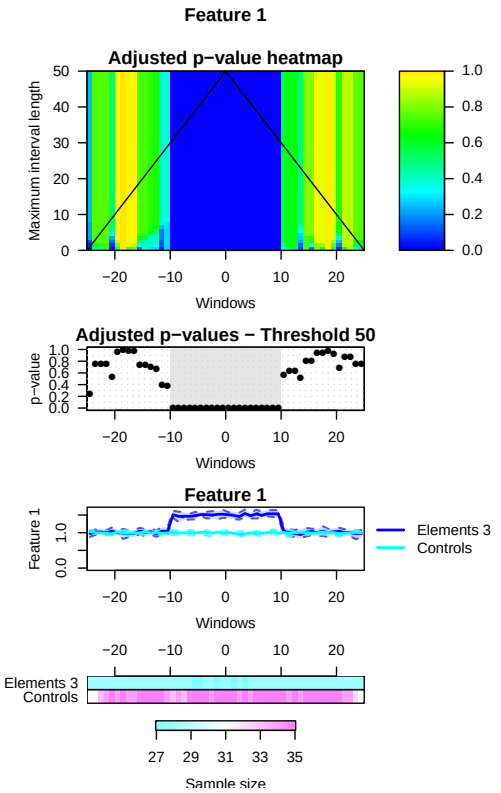
2 *IWTomics* test results and workflow



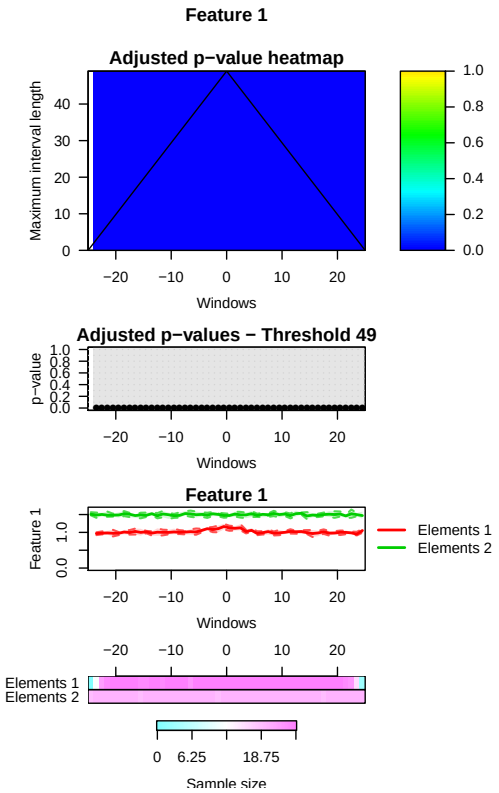
(a)



(b)



(c)



(d)

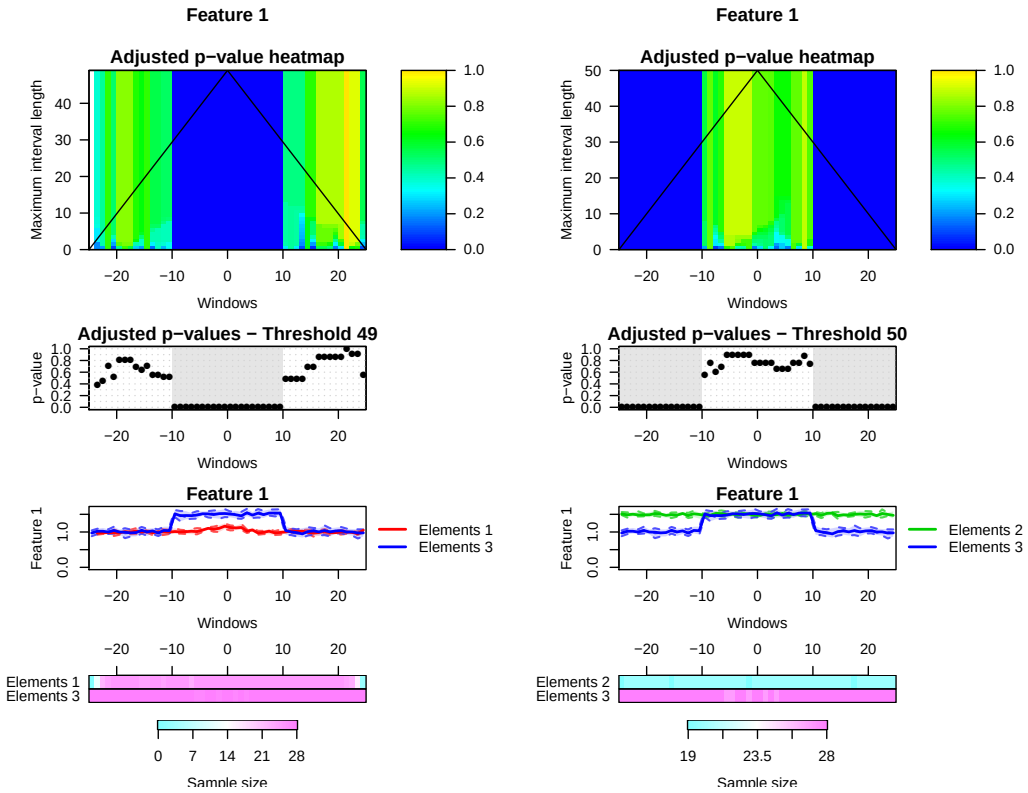


Figure S1: Graphical representation of IWT results on the simulated dataset shown in Fig. 1. (a) Elements 1 vs Controls; (b) Elements 2 vs Controls; (c) Elements 3 vs Controls; (d) Elements 1 vs Elements 2; (e) Elements 1 vs Elements 3; (f) Elements 2 vs Elements 3. In each panel, the heatmap at the top shows the adjusted p-value curve for each possible scale (the bottom row shows scale 1, i.e. the unadjusted p-value curve; the top row shows the maximum scale possible). The middle plot shows the adjusted p-value curve at the selected scale threshold, with significant locations (p-value< 0.05) highlighted in gray. The bottom plot shows the pointwise boxplot of the curves in the two groups contrasted by the test, and the sample size corresponding to each window in each group.

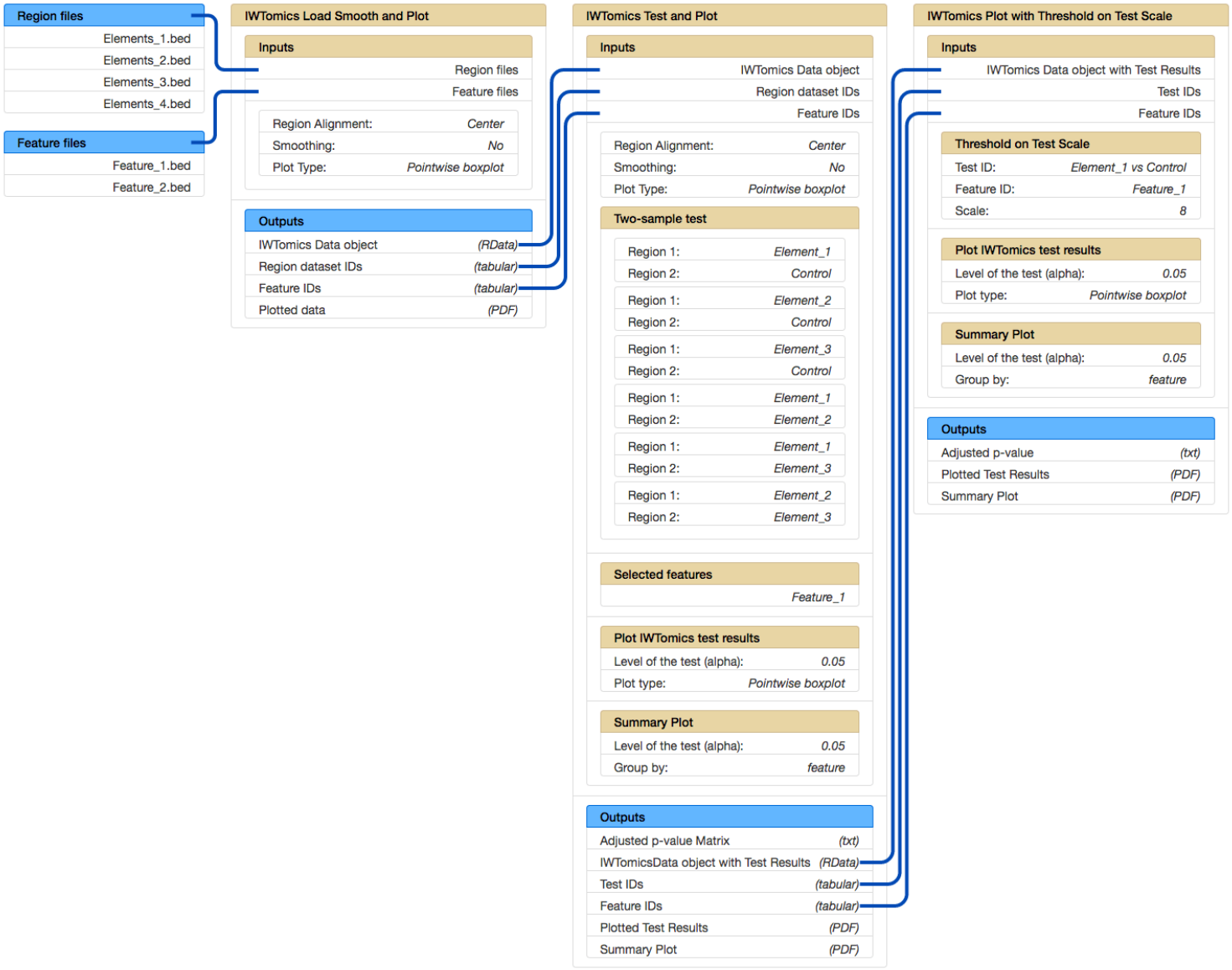


Figure S2: Workflow of the Galaxy tool, on the simulated dataset shown in Figs. 1 and S1.

3 Comparing *IWTomics* with other methods

We demonstrate the effectiveness of *IWTomics* and its advantages over alternative approaches utilizing both real biological data and simulated data examples.

3.1 Real data example: recombination hotspots in fixed ETn

We consider a real biological example concerning recombination hotspots in the flanking regions of fixed ETns (elements of the Early Transposon family of active Endogenous Retroviruses in mouse). The dataset comprises a total of 2438 64-kb regions in the mouse genome; 1296 surrounding ETn elements (32-kb upstream and 32-kb downstream of each element) and 1142 controls (64-kb regions). Recombination hotspots are measured in 1-kb windows inside each region (see Fig. S3). This dataset is a subset of the one collected and analyzed in Campos-Sánchez *et al.* (2016), and it is provided within the R package *IWTomics*.

To investigate whether recombination plays a role in ETn fixation in the mouse genome, we contrast measurements of recombination hotspots in the flanks of fixed ETn and in control regions. We do this with four different approaches; namely:

1. *IWTomics*; this computes an adjusted p-value curve taking into consideration contiguity of measurements in adjacent windows (see Suppl. Section 1);

2. a *global t-test*; this considers the average signal across the 64 windows in each region, and does not involve a p-value correction;

3. a collection of 64 *local t-tests*; this considers separately the signals in the 64 window in each region, and does involve a p-value correction (see below);

4. a *graphical contrasting of average profiles*, which can be enriched for inferential purposes (again, see below).

For the multiple testing correction in (3) we consider *comb-p* (Pedersen *et al.*, 2012), a state-of-the-art recent method whose underpinning is similar to that of *IWTomics*. Thus, performance of (3) with *comb-p* should be the most competitive with *IWTomics*. *Comb-p* corrects accounting for spatial correlation between the p-values. Notably though, and unlike *IWTomics*, it requires fixing a tuning parameter – the maximum lag for computing autocorrelations.

For the approach in (4), we employ one of very many existing software packages that allow the user to plot ChIP-seq read counts average profiles (see e.g. Stempor and Ahringer, 2016; Anders *et al.*, 2015); namely the R package *ChIP-seeker* (Yu *et al.*, 2015). We select *ChIP-seeker* because it is the most adaptable to data other than read coverage (*IWTomics* can be applied to any kind of data) and because, unlike other packages, it has an inferential component; it provides pointwise bootstrap confidence intervals. We note that in *ChIP-seeker* the signals for each window within each region are normalized, and this is done separately for the two groups of regions (more specifically, since the data are meant to be read counts, frequencies are computed dividing the sum of measurements for a given window by the total sum of measurements over all windows). Since *ChIP-seeker* represents differences between such normalized averages in each window, a constant difference along the whole region would disappear because of the normalization. Table S1 summarizes the characteristics of the different approaches considered for comparison, highlighting some of the advantages of *IWTomics*; its use of permutations and its ability to work with different test statistics make it more flexible, its technique to correct for multiple testing is very effective because it leverages contiguity of measurements, it is capable of detecting both significant locations and significant scales and, because of its functional data underpinning, it can handle NAs (missing values).

We can observe in Fig. S3 that recombination hotspots content shows a small difference in mean between the flanks of fixed ETn and control regions. This difference is localized near the

Method	Type of inference	Test statistic	Multiple testing correction	Handle NA	Detect location	Detect scale
IWTomics	Permutation	Multiple statistics	Yes (adjacent windows)	Yes	Yes	Yes
Global t-test	Parametric	Mean	NA	Yes	No	No
Local t-test with *comb-p*	Parametric	Mean	Yes (correlated windows)	Yes	Yes	No
ChIP-seeker	Bootstrap	Normalized mean	No	No	Yes	No

Table S1: Comparison of *IWTomics* with other existing approaches.

ETn insertion site (i.e. the center of the 64-kb region). *IWTomics* (with mean difference test statistic) captures this very localized effect in the subregion between 5 kb upstream and 1 kb downstream of the insertion site, and suggests a scale of 8-kb (see Fig. S4). The global t-test on the average signal in the 64-kb region fails to capture a significant difference between the two groups (p-value 0.32). The local t-test with *comb-p* correction produces significant results around the insertion site, but it identifies a larger subregion than *IWTomics*. More importantly, the subregion width strongly depends on the choice for the maximum lag to be employed for the autocorrelation computation (Fig. S5). For example, an extreme maximum lag of 50 kb (the largest viable on the data; for larger lags we do not have enough windows to estimate the

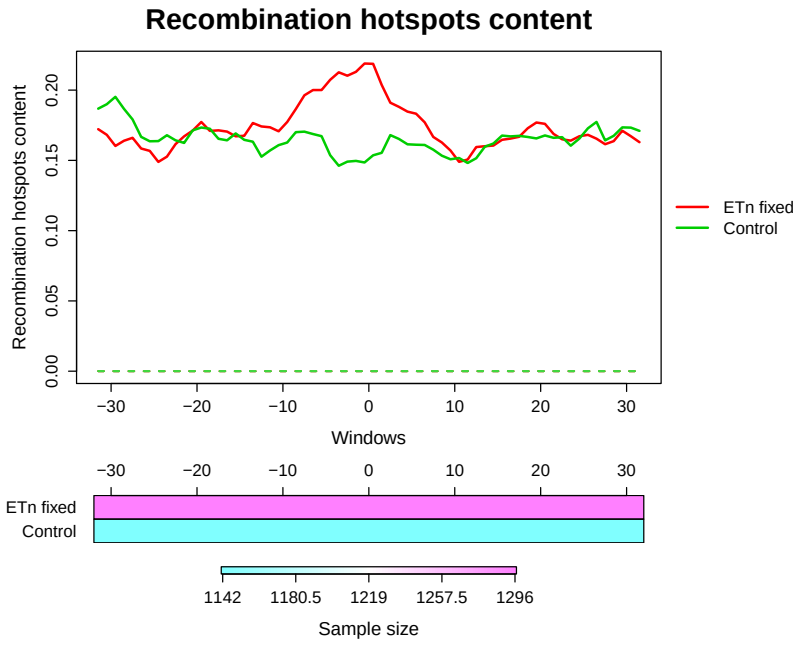


Figure S3: Plot of recombination hotspots content in the flanking regions of fixed ETns (red) and in control regions (green). Solid lines represent the average curves for the two types of regions (average signals in each 1-kb window), while dashed lines represent pointwise quartiles (all identically equal to zero for these data). The heatmap at the bottom shows the sample size in each position.

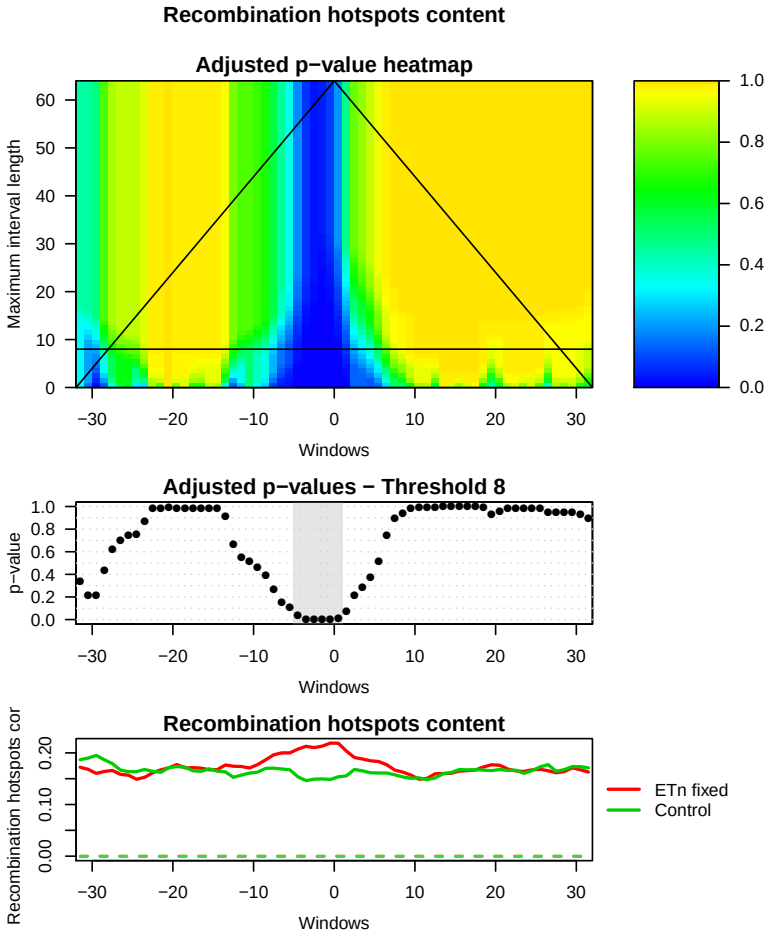


Figure S4: Graphical representation of IWT results for the test ETn fixes versus control. See caption of Fig. S1.

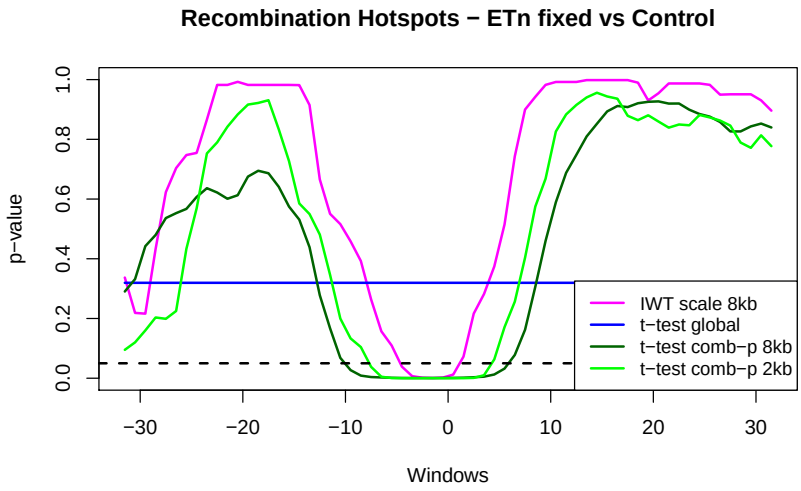


Figure S5: Comparison of p-values obtained with different methods. The black dashed line indicates a 0.05 significance level.

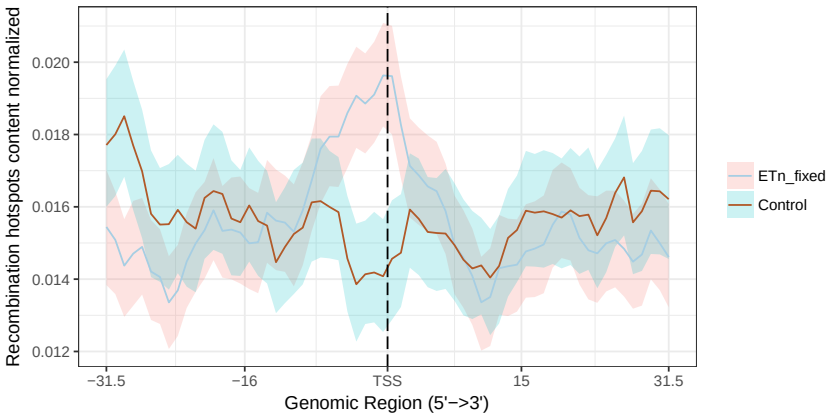
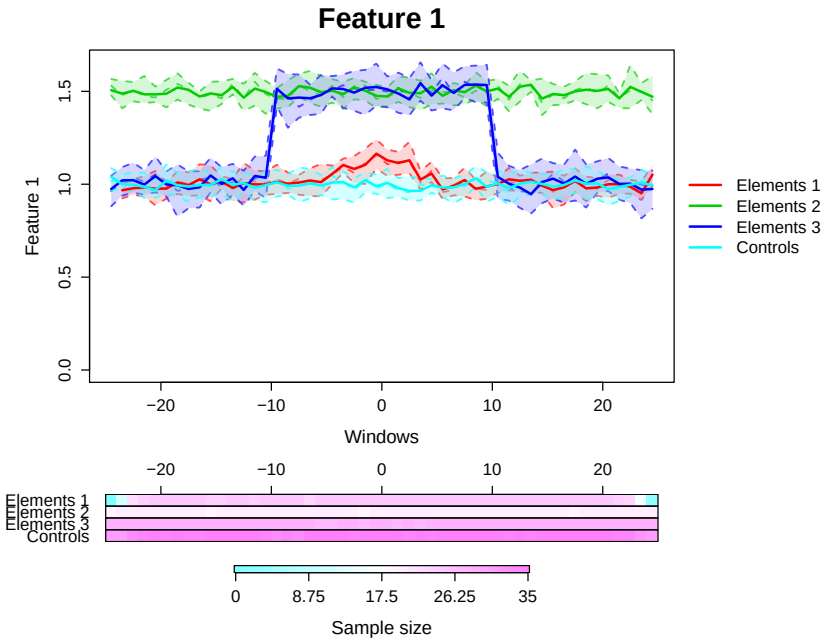


Figure S6: Average profile plot obtained with *ChIP-seeker*. The semi-transparent bands indicate the 95% pointwise bootstrap confidence interval for the average profile of each group.

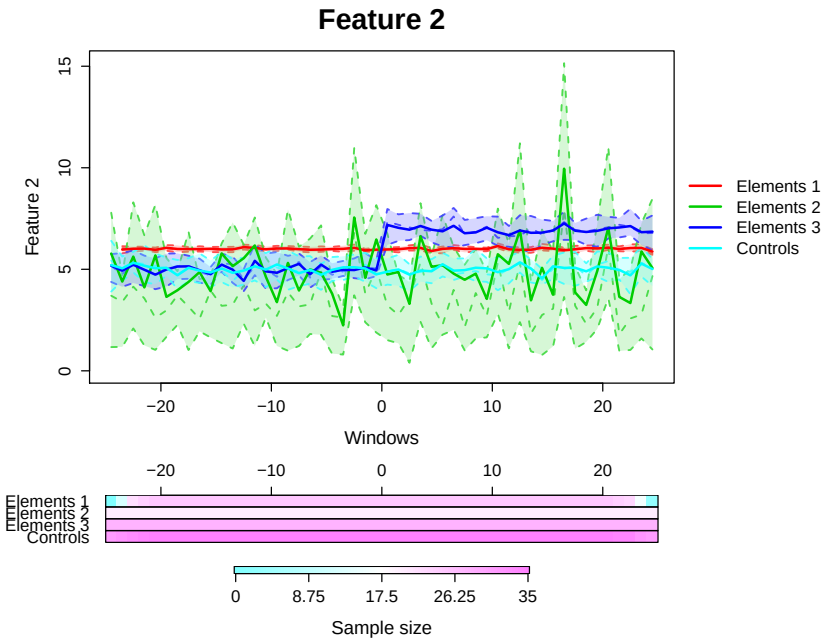
autocorrelation) produces p-values< 0.05 for the whole 64-kb region. A maximum lag of 8 kb (the one at which the autocorrelation comes close to 0 for the first time) finds a significant difference in the subregion between 10 kb upstream and 6 kb downstream of the insertion site. The results more similar to those of *IWTomics* are obtained using a maximum lag of 2 kb, which leads to the identification of the subregion between 8 kb upstream and 4 kb downstream of the insertion site. Contrasting average profiles with *ChIP-seeker* (Fig. S6) indicates a significant difference in the central part of the region, and also in the window 30 kb upstream of the insertion site. This is is due to the fact that pointwise bootstrap confidence intervals produced by *ChIP-seeker* do not correct for multiple testing, hence the probability of false positive is higher than the nominal one.

3.2 Simulated data example

To further compare the performances of *IWTomics* and of the other approaches listed in Table S1, we consider a simulated dataset provided with the *IWTomics* package ("regionsFeatures_center" data) and used in Figs. 1 and S1. This dataset comprises four groups of regions ("Elements 1", "Elements 2", "Elements 3" and "Controls") and two features ("Feature 1" and "Feature 2"). The regions have different lengths, up to 100 kb, and they are aligned at their center. The features resolution is 2 kb (hence the maximum number of windows is 50), and there are some missing data (NA). Part of this dataset has been used to produce Figs. 1 and S1; Fig. S7 shows the dataset in its entirety. Since we know the ground truth (in terms of difference in distributions), we evaluate the methods by computing the number of true and false positive windows identified by each. For *IWTomics*, we consider two specifications of the test statistics; mean difference and multiple quantile difference (with quantiles 0.25, 0.50 and 0.75). Corresponding to each specification, we also report the scale identified by the p-value heatmap. For the *local t-test* with *comb-p* adjustment, we consider two choices for the maximum lag; the one suggested by the autocorrelation plot, as well as the best one (i.e. the maximum lag that leads to best true and false positives numbers). For *ChIP-seeker*, since the tool does not handle missing data (NA), we need to make an arbitrary choice and set all NAs to 0 (i.e. the baseline level for read coverage data). Table S2 shows the results of this comparison and indicates that *IWTomics* usually performs better than the other approaches, being both highly sensitive and highly specific.



(a)



(b)

Figure S7: Pointwise boxlot of (a) “Feature 1” and (b) “Feature 2” in the four groups of regions “Elements 1”, “Elements 2”, “Elements 3” and “Controls”.

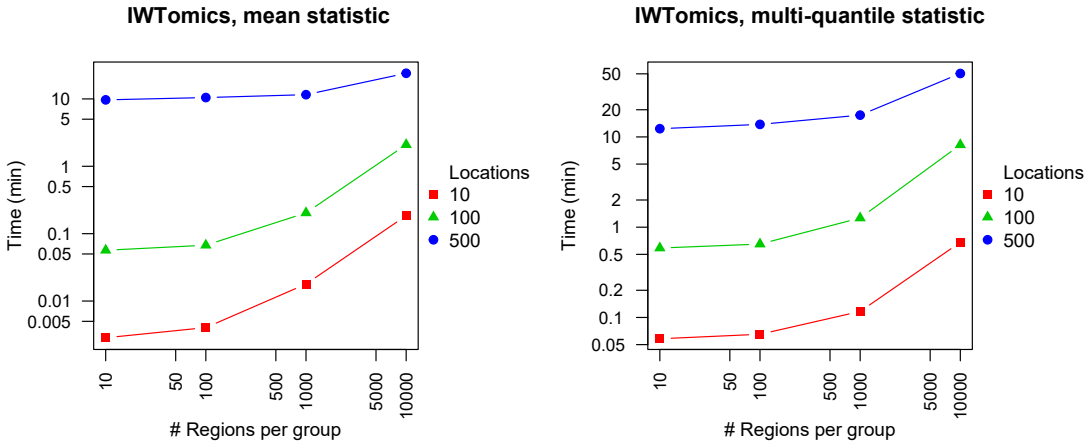
Method	Feature 1			Feature 2		
	Element 1 vs Control	Element 2 vs Control	Element 3 vs Control	Element 1 vs Control	Element 2 vs Control	Element 3 vs Control
IWTomics mean	scale 16kb 7/8 (0/41)	scale 100kb 50/50 (0/0)	scale 100kb 20/20 (0/30)	scale 98kb 49/49 (0/0)	scale 18kb 3/50 (0/0)	scale 100kb 25/25 (0/25)
IWTomics quantiles	scale 16kb 6/8 (1/41)	scale 100kb 50/50 (0/0)	scale 100kb 20/20 (0/30)	scale 98kb 49/49 (0/0)	scale 100kb 50/50 (0/0)	scale 100kb 25/25 (0/25)
Global t-test	8/8 (41/41)	50/50 (0/0)	20/20 (30/30)	49/49 (0/0)	0/50 (0/0)	25/25 (25/25)
Local t-test with *comb-p* autocor	lag 8kb 8/8 (5/41)	lag 8kb 50/50 (0/0)	lag 14kb 20/20 (14/30)	lag 6kb 49/49 (0/0)	lag 6kb 17/50 (0/0)	lag 32kb 25/25 (12/25)
Local t-test with *comb-p* best	lag 2kb 8/8 (3/41)	lag 8kb 50/50 (0/0)	lag 2kb 20/20 (4/30)	lag 6kb 49/49 (0/0)	lag 70kb 50/50 (0/0)	lag 2kb 25/25 (1/25)
ChIP-seeker	8/8 (7/41)	0/50 (0/0)	20/20 (25/30)	11/49 (0/0)	5/50 (0/0)	19/25 (17/25)

Table S2: Comparison of *IWTomics* with other existing approaches. Each cell reports the ratio of true positive windows, and in parentheses the ratio of false positive windows (note that the total number of windows is 50, corresponding to 100kb, for "Elements 2" "Elements 3", while "Elements 1" has 49 windows, corresponding to 98kb).

4 *IWTomics* runtime

IWTomics runtime depends on several characteristics of the dataset and of the test performed. First, it depends on the feature measurement resolution, and thus on the number of locations/scales (K) considered in the test. Computation of the unadjusted p-value curve (scale 1) is linear in K, while the adjustment procedure is in the order of K^2. Second, runtime increases with the sample size, i.e. with the number of regions available in the different groups. Third, it depends on the test statistic employed. In particular, *mean difference* is the fastest among the provided test statistics, while *quantile difference* is the slowest because its computation requires a sorting step. Finally, runtime is linear in the number of random permutations (B) employed to approximate the p-values, and in the number of tests performed.

In order to provide an indication about *IWTomics* runtime requirements, we present a simulation analysis performed on a 64-bit Debian GNU/Linux 7 OS running on a 2.3 GHz processor with 8 cores and 16 GB RAM. The simulation considers a test between two groups of regions, with equal sample size ranging between 10 and 10000 regions, and number of locations/scales $K = 10, 100$ and 500. A total of $B = 1000$ random permutations are employed to compute p-values (default choice in the package; this provides a p-value resolution of 0.001). Fig. S8 shows simulation results concerning (a) *mean difference* and (b) *quantile difference* (with 0.25, 0.50 and 0.75 quantiles) test statistics. Points represent the average runtime across 5 independent



(a) (b)

Figure S8: *IWTomics* runtime analysis on simulated data. (a) *Mean difference* test statistic; (b) *Quantile difference* test statistic (quantiles of order 0.25, 0.50 and 0.75).

runs (interestingly, the 5 runtimes were always very close to each other, hence we provide only average values).

Real applications of *IWTomics* such as the ones discussed in Section 3 typically involve groups of a few hundreds to a few thousands of genomic regions, with features measured on $50 - 100$ windows. In these cases, a test based on the *mean difference* statistic takes less than half minute to run, while a test based on the *quantile difference* statistic is in the order of 1 minute. Importantly, *IWTomics* has a reasonable runtime even when 10000 genomic regions are considered in each of the two groups, and the test is performed on 500 locations/scales.

5 Supplementary References

Examples of prior literature employing window or profile-based approaches to study and contrast genomic features:

Hellmann, I. *et al.* (2005). Why do human diversity levels vary at a megabase scale? *Genome Res.*, **15**(9), 1222–1231

Kvikstad, E. M. *et al.* (2007). A macaque's-eye view of human insertions and deletions: differences in mechanisms. *PLOS Comput. Biol.*, **3**(9), e176

Yu, G., Wang, L.-G., and He, Q.-Y. (2015). ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics*, **31**(14), 2382–2383

Ernst, J. and Kellis, M. (2010). Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat. Biotechnol.*, **28**(8), 817–825

Dale, R. K. *et al.* (2014). metaseq: a Python package for integrative genome-wide analysis reveals relationships between chromatin insulators and associated nuclear mRNA. *Nucleic Acids Res.*, **42**(14), 9158–9170

A method for combining p-values in adjacent windows accounting for spatial correlation along the genome:

Pedersen, B. S., Schwartz, D. A., Yang, I. V., and Kechris, K. J. (2012). Comb-p: software for combining, analyzing, grouping and correcting spatially correlated P-values. *Bioinformatics*, **28**(22), 2986–2988

IWTomics comprises the S4 data class *IWTomicsData*, which extends the commonly used Bioconductor class *GRangesList*:

Lawrence, M. *et al.* (2013). Software for computing and annotating genomic ranges. *PLOS Comput. Biol.*, **9**(8), e1003118

Prior applications of FDA to "Omics", which concerned GWAS with functional outcomes and ChIP-seq analyses:

Reimherr, M. and Nicolae, D. (2014). A functional data analysis approach for genetic association studies. *Ann. Appl. Stat.*, **8**(1), 406–429

Cremona, M. A. *et al.* (2015). Peak shape clustering reveals biological insights. *BMC Bioinform.*, **16**(1), 349

Parodi, A. *et al.* (2017). FunChIP: an R/Bioconductor package for functional classification of ChIP-seq shapes. *Bioinformatics*, page btx201

Madrigal, P. (2016). fCCAC: functional canonical correlation analysis to evaluate covariance between nucleic acid sequencing datasets. *Bioinformatics*, page btw724

Mateos, J. L., Madrigal, P., *et al.* (2015). Combinatorial activities of short vegetative phase and flowering locus c define distinct modes of flowering regulation in arabidopsis. *Genome Biol.*, **16**(1), 31

The ETn dataset employed here has been collected, pre-processed and analyzed in:

Campos-Sánchez, R., Cremona, M. A., *et al.* (2016). Integration and fixation preferences of human and mouse endogenous retroviruses uncovered with functional data analysis. *PLOS Comput. Biol.*, **12**(6), 1–41