BMC Bioinformatics

METHODOLOGY ARTICLE

Open Access

Detecting differentially methylated regions using a fast wavelet-based approach to functional association analysis

William R. P. Denault[1,2,3*] and Astanand Jugessur[1,2,3]

*Correspondence:
william.denault@fhi.no
[1] Department of Genetics and Bioinformatics, Norwegian Institute of Public Health, Oslo, Norway
Full list of author information is available at the end of the article

Abstract

Background: We present here a computational shortcut to improve a powerful wavelet-based method by Shim and Stephens (Ann Appl Stat 9(2):665–686, 2015. https://doi.org/10.1214/14-AOAS776) called WaveQTL that was originally designed to identify DNase I hypersensitivity quantitative trait loci (dsQTL).

Results: WaveQTL relies on permutations to evaluate the significance of an association. We applied a recent method by Zhou and Guan (J Am Stat Assoc 113(523):1362–1371, 2017. https://doi.org/10.1080/01621459.2017.1328361) to boost computational speed, which involves calculating the distribution of Bayes factors and estimating the significance of an association by simulations rather than permutations. We called this simulation-based approach "fast functional wavelet" (FFW), and tested it on a publicly available DNA methylation (DNAm) dataset on colorectal cancer. The simulations confirmed a substantial gain in computational speed compared to the permutation-based approach in WaveQTL. Furthermore, we show that FFW controls the type I error satisfactorily and has good power for detecting differentially methylated regions.

Conclusions: Our approach has broad utility and can be applied to detect associations between different types of functions and phenotypes. As more and more DNAm datasets are being made available through public repositories, an attractive application of FFW would be to re-analyze these data and identify associations that might have been missed by previous efforts. The full R package for FFW is freely available at GitHub https://github.com/william-denault/ffw.

Keywords: Wavelets, DNA methylation, EWAS, Association analysis, Epigenetics

Background

Despite the recent surge of interest in functional association analysis of various types of high-dimensional data, e.g., those from biomechanical research [1], quantitative trait loci (QTL) mapping [2], genome-wide association studies (GWASes) [3], and epigenome-wide association studies (EWASes) [4, 5], the majority of genome-wide screenings are still largely based on testing one SNP or one CpG at a time (single-point analysis). Single-point analyses are limited because they incur a substantial

multiple-testing burden and ignore the genomic context of the association. Bump-hunting methods [4, 6] can, to some extent, alleviate these issues by detecting a predetermined difference in the spatially ordered variables (e.g., SNPs or CpGs) that correlate with the trait or disease under scrutiny. However, as the name implies, bump-hunting methods can only analyze a single bump at a time, failing to consider the combined effect of multiple bumps in a given region. Besides the lack of power [5], bump-hunting methods also rely on a re-sampling procedure, such as bootstrapping or permutation, which makes them computationally more intensive (for details, see [4]).

To help address these methodological shortcomings, we developed a wavelet-based method to enable the detection of more complex signals than those present in a single bump. Although several methods are now available for functional association analysis based on wavelets [5, 7–9], they do not scale well in terms of computational time when the sample size (n) or the number of regions (R) becomes exceedingly large (e.g., when $n \approx 1000$ or $R > 1000$). As wavelet coefficients are not independent, Lee and Morris [5] and Ma and Soriano [9] proposed searching for associations between a trait and a function by using wavelet regression that takes into account the dependencies between the wavelet coefficients. However, this requires the use of a re-sampling procedure such as Markov Chain Monte Carlo (MCMC) (exemplified in [5, 7]) or the computation of complex analytical posterior distributions as in Ma and Soriano [9]. To address these issues, Shim and Stephens [8] proposed simplifying the modeling by omitting the dependencies and using a likelihood ratio test to search for associations between a trait and a function. This simplification is, however, still limited because of the need for permutations to evaluate the significance of the likelihood ratio.

In our current approach, which we call "Fast Functional Wavelet" (FFW), we combine the framework of Shim and Stephens [8] with recent results on the theoretical null distribution of Bayes factors by Zhou and Guan [10]. Our approach allows a fast emulation of the test statistic in Shim and Stephens [8] and reduces the computational time considerably. The main difference between FFW and WaveQTL is that FFW requires regressing the trait of interest on the wavelet coefficients, regardless of the application. Hence, the design matrix for association is kept constant across all the screened regions. This simple modification enables the null distribution of the test statistic to be simulated using only a χ^2_1 distribution, thereby circumventing the need for extensive permutations to assess the significance of a region. By keeping the design matrix for association constant across the screened regions, and using the results of Zhou and Guan [10], we can show that the same distribution can be used to emulate the null distribution of each regional test. This null distribution depends on a single parameter that is easily computed. Keeping the design matrix constant across the screened regions can lead to a reverse regression, resulting in shrunken estimates and a potential loss of power (see Fuller, Chapter 1 [11]). Besides making the null distribution easier to emulate, reverse regression also allows the analysis of a wider variety of traits (continuous, binary, or count).

Although the focus of the current paper is on DNAm data, we describe FFW more broadly to highlight its utility for other types of high-dimensional data, such as those

from a GWAS [3], dsQTL analysis [2], and functional data from biomechanical research [1].

Description of our wavelet-based approach

There are different types of wavelets, and readers interested in a more comprehensive introduction to wavelets are referred to Nason [12]. In our application here (FFW), we only consider the simplest type of wavelet—the Haar wavelet.

Processing

For simplicity, we consider genetic regions composed of evenly-spaced variables of length 2^J. As the assumption of evenly-spaced variables rarely holds in practice, we use the approach of Kovac and Silverman [13] to handle this limitation in our software implementation. Their approach mainly consists of using an interpolation on the observed data to obtain evenly-spaced variables of length 2^J.

A given region is defined as a compound of $X_1, \ldots X_{2^J}$, with 2^J spatially-ordered variables. Suppose we observe these 2^J variables for n individuals; we then denote $X_{i,k}$ as the kth observation of the ith variable. The wavelet coefficients for the individual k are defined as follows. For wavelet coefficients corresponding to the highest scale or resolution (i.e., J), and for $i \in [\![1, 2^{J-1}]\!]$

$$\tilde{X}_{J,i,k} = X_{2i,k} - X_{2i-1,k}$$

These wavelet coefficients correspond to local differences between adjacent variables. For a lower scale (i.e., $j < J$), the corresponding wavelet coefficients are computed as follows:

$$\tilde{X}_{j,i,k} = \check{X}_{j+1,2i,k} - \check{X}_{j+1,2i-1,k}$$

where $\check{X}_{j,2i,k}$, is defined as:

$$\check{X}_{j,2i,k} = \begin{cases} \forall i \in [\![1, 2^{J-1}]\!], X_{2i,k} + X_{2i-1,k}, & \text{if } j = J \\ \forall i \in [\![1, 2^{j-1}]\!], \check{X}_{j+1,2i,k} + \check{X}_{j+1,2i-1,k}, & \text{if } 1 \leq j < J \end{cases} \quad (1)$$

$\check{X}_{j,2i,k}$ correspond to the scaled average of the adjacent variables for individual k (for further details, see Nason [14]). Finally, the wavelet coefficients for the lowest scale (i.e., 0) are computed as follows:

$$\tilde{X}_{0,1,k} = \sum_{i=1}^{2^J} X_{i,k}$$

The procedure described above is known as Mallat's pyramid algorithm for signal processing [15]. To ease comprehension, we denote $\tilde{X}_{jl}$ as the random variable representing the wavelet coefficient at the scale j, with $1 \leq j \leq J$, and at the location l, with $1 \leq l \leq 2^{j-1}$.

Modeling

We first summarize the work of Shim and Stephens [8] before presenting our main methodological contributions in the "Significance of a region" section further below.

To identify associations between a region and a phenotype (denoted as Φ hereafter), we assess whether specific scales are associated with Φ at different locations. Let π be a vector of length J, where $\forall j \in [0:J], \pi_j \in [0,1]$ and π_j represents the proportion of wavelet coefficients at scale j associated with Φ. To assess the significance of a given genetic region, we test the following hypothesis:

$$H_0 : \pi = (0,\ldots,0) vs H_1 : \exists j \in [0:J], \pi_j \neq 0 \tag{2}$$

In the next sections, we describe the test statistic (likelihood ratio), how its different components are computed, and how its significance is tested.

Bayes factors

To test for associations between Φ and the wavelet coefficient $\tilde{X}_{jl}$ for a given region, we use the Normal-Inverse-Gamma (NIG) prior to perform a regression between each wavelet coefficient and Φ. Note that our framework easily allows confounders to be incorporated into the regression models. We quantile-transform each wavelet coefficient across the individuals to reduce the proportion of spurious associations due to distribution-related issues.

The association models for each scale and location are defined as follows:

$$\begin{aligned} M_0 &: \tilde{X}_{jl} = \beta_{jl,0} + \beta_{jl,C} C + \epsilon \\ M_1 &: \tilde{X}_{jl} = \beta_{jl,0} + \beta_{jl,1}\Phi + \beta_{jl,C} C + \epsilon \end{aligned} \tag{3}$$

where C is a matrix of dimension $c \times n$, $\beta_{jl,C}$ is a matrix of dimension $1 \times c$ and $\epsilon \sim N(0,\sigma^2)$, where σ^2 is unknown. Next, we compute the Bayes factors of the wavelet regression jl using the closed form provided by Zhou and Guan [10] for the NIG prior.

Ratio statistic

Our goal is to assess the significance of the vector $\pi = (\pi_0,\ldots,\pi_j,\ldots,\pi_J)$, where π_j represents the proportion of wavelet coefficients at scale j associated with Φ, and $\tilde{X}$ is the wavelet representation of the individual functions. To test the significance of π, we construct a test statistic by computing the following likelihood ratio:

$$\Lambda(\pi, \tilde{X}, \Phi) = \frac{p(\tilde{X}|\pi, \Phi)}{p(\tilde{X}|\pi \equiv 0, \Phi)} \tag{4}$$

Following the approach of Shim and Stephens [8], we denote γ_{jl} as the random variable with support $\{0,1\}$. Thus, $\gamma_{jl} = 1$ if the wavelet coefficient $\tilde{X}_{jl}$ is associated with Φ, and 0 if not. We consider π as a hyperparameter of γ_{jl}; i.e.,

$$p(\gamma_{jl} = 1|\pi) = \pi_j \tag{5}$$

Shim and Stephens [8] assume independence between the wavelet coefficients. However, this may not to hold in practice [16]. Under the assumption of independence of the wavelet coefficients, we can rewrite 4 as:

$$\Lambda(\pi, \tilde{x}, \Phi) = \prod_{j,l} \frac{p(\tilde{x}_{jl}|\pi_j, \Phi)}{p(\tilde{x}_{jl}|\pi_j = 0, \Phi)} \tag{6}$$

$$= \prod_{j,l} \frac{\pi_j p(\tilde{x}|\gamma_{jl} = 1, \Phi) + (1 - \pi_j) p(\tilde{x}|\gamma_{jl} = 0, \Phi)}{p(\tilde{x}|\gamma_{jl} = 0, \Phi)} \tag{7}$$

We denote $BF_{jl}(\tilde{x}, \Phi) = \frac{p(\tilde{x}_{jl}|\gamma_{jl}=1,\Phi)}{p(\tilde{x}_{jl}|\gamma_{jl}=0,\Phi)}$ as the Bayes factor of the association between the wavelet coefficient at scale s and location l. Using this notation, we can rewrite 7 as:

$$\Lambda(\pi, \tilde{x}, \Phi) = \prod_{j,l} \left[\pi_j BF_{jl} + (1 - \pi_j)\right] \tag{8}$$

We then compute the likelihood ratio statistic by maximizing the lambda statistics over π and estimating $\hat{\pi}$ using the EM algorithm.

$$\hat{\Lambda}(\tilde{x}, \Phi) = max_{\pi \in [0,1]^J} \Lambda(\pi, \tilde{x}, \Phi) \tag{9}$$

Significance of a region

As the distribution of Λ is unknown, we simulate Λ under H_0 by simulating BF_{jl} under H_0. Recently, Zhou and Guan [10] showed that, under H_0 and a wide spectrum of priors, the Bayes factors (including the NIG prior) follow a specific distribution for a Gaussian model. More precisely,

$$2log(BF) = \lambda_1 Q_1 + log(1 - \lambda_1) + \epsilon \tag{10}$$

where Q_1 is a non-central chi-squared random variable with $df = 1$, and $\epsilon = O(1)$ and its non-centrality parameter has a closed-form. The parameter λ_1 is the largest eigenvalue of $X\left(X^t X + V_b^{-1}\right)^{-1} X^t$, where X is the design matrix. Specifically, $X = \left(\mathbf{1}, Y, C^t\right)$ and $V_b = diag(\sigma_b^2)$ (σ_b is the prior effect size of the NIG prior for the intercept and the covariates). By keeping the design matrix constant across the regions, λ_1 stays the same for all the regions and only needs to be computed once. The non-centrality parameter is region-dependent in general, but it is exactly zero when the null hypothesis of the Bayes factor is $\beta_{jl,1} = 0$. Zhou and Guan [10] showed that, for $df = 1$, Q_1 is asymptotically equal to the likelihood ratio test statistic for Gaussian linear models. In other words, Q_1 is equal to a simple chi-squared statistic with one degree of freedom. Therefore, we use the approximation in Eq. (11) for the distribution of the Bayes factors. Note also that Zhou and Guan [10] showed that this approximation is exact when using a Normal prior.

$$2log(BF) \approx \lambda_1 \chi_1^2 + log(1 - \lambda_1) \tag{11}$$

By using this approximation, it is only necessary to compute a single parameter for all the regions. We can then perform M independent simulations of the vector of Bayes

factors under H_0. This corresponds to M vectors of length 2^J, corresponding to the number of wavelet coefficients. Next, for each simulated vector of Bayes factors BF_m, we compute the simulated likelihood ratio $\hat{\Lambda}_m = max_{\pi \in [0,1]^J} \Lambda(\pi, BF_m)$ using the procedure described above. Monte Carlo methods for p value estimation can then be applied to the set of observed statistics.

Simulations and application

We performed a set of simulations to evaluate the gain in computational time with FFW and to assess the significance of the test statistic. We also evaluated the statistical power of FFW using a realistically simulated dataset. Lastly, we ran a sensitivity analysis for the priors to assess the sensitivity of FFW according to the choice of prior. All the simulations were performed on an ordinary laptop, equipped with an Intel(R) i7-700HQ 2.80 GHz processor and 8 GB of RAM.

Gain in computational time

We performed separate EWASes using FFW and WaveQTL, and report the run time for different sample sizes. Figure 1 illustrates the substantial gain in computational time with FFW. The green and orange curves represent the total time it took to perform an EWAS based on DNAm data generated on the Illumina 450K platform (using the same pre-processing steps as in the "Power and prior sensitivity analysis" section).

The data used to estimate the computational time were generated as follows. First, we simulate each individual's DNAm profile using independent and identically distributed (iid) uniform random variables on [0, 1]. More precisely, we simulate each individual's DNAm as being generated by the Illumina 450K array, by simulating the value of each probe on the array using a random variable on [0, 1]. Next, we apply the same

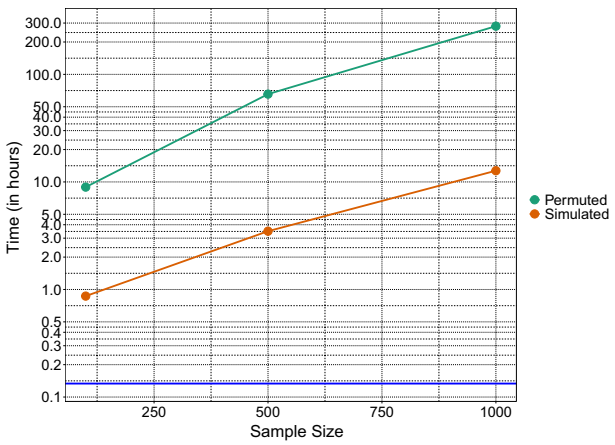

Fig. 1 Comparison of the computational time under H_0 when performing an EWAS with FFW and WaveQTL for different sample sizes. The y-axis has been log-transformed. The green curve indicates the computational time using the permutation procedure described by Besag and Clifford [29], which is based on a maximum of 10 permutations being larger than the observed test statistics (the set up is according to Shim and Stephens [8]). The orange curve indicates the computational time using FFW. The blue line indicates the computational time of the simulations required in FFW, which remains constant with increasing sample size

pre-processing steps as in the "Power and prior sensitivity analysis" section, resulting in 4731 regions to test for association.

Finally, in the "Computational cost" section of the Appendix, we derive the theoretical computational cost for WaveQTL and FFW.

Type I error

We estimated the type I error for four distinct scenarios and performed simulations under the assumption of no association, using the test functions *block*, *bump*, *heaviSine*, and *doppler* as previously described in Donoho and Johnstone [17]. The different functions are illustrated in Fig. 2 (adapted from Donoho and Johnstone [17]).

For each simulation s, we propose the following model of association. For each test function T, we perform the following simulation, and using T, we generate a population of observed functions as follows:

$$f_k(x) = a_k \times T(x) + \epsilon$$

where a_k is the individual amplitude of the test function, with $a \sim N(0,1)$ and $\epsilon \sim N(0,1)$. We denote F as the set of observed functions. We then generate $Y_k \sim N(0,1)$, which represents a continuous trait not associated with the considered test function. Next, we wavelet-transform each individual function f_k and quantile-transform each wavelet coefficient in the population. We then compute the likelihood ratio $\hat{\Lambda}(\tilde{F}, Y)_s$, as well as λ_{1_s}, which is the largest eigenvalue of $X\left(X^tX + V_b^{-1}\right)^{-1}X^t$, where $X = (\mathbf{1}, Y)$ and $V_b = diag(\sigma_b^2)$ (σ_b is the prior effect size of the NIG prior for the intercept and the covariate). We simulate $\hat{\Lambda}(\tilde{F}, Y)_s$ $M = 10^6$ times using

$$2log(BF) \approx \lambda_{1_s}\chi_1^2 + log(1 - \lambda_{1_s}) \tag{12}$$

These simulations are denoted as $\hat{\Lambda}(\tilde{F}, Y)_s^m$. We compute the Monte Carlo p value as

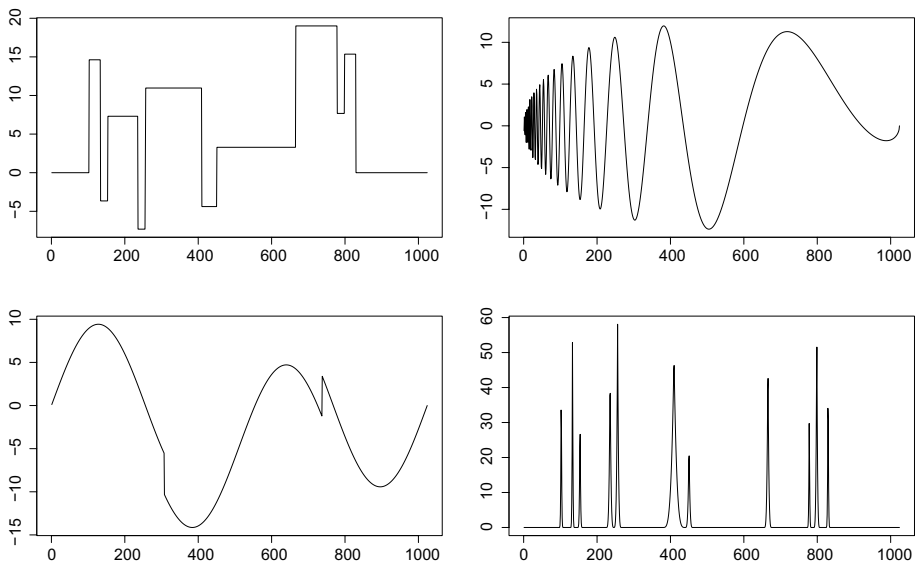

Fig. 2 Test functions evaluated for type I error. Top left: Block; top right: Doppler; bottom left: HeaviSine; and bottom right: Bump

$$\hat{p}_s = \frac{Card\left(m, \hat{\Lambda}(\tilde{F}, Y)_s \geq \hat{\Lambda}(\tilde{F}, Y)_s^m\right) + 1}{M + 1} \tag{13}$$

This procedure is repeated 75,000 times for each sample size and type of test function (*block*, *bump*, *heaviSine*, and *doppler*). Table 1 summarizes the estimated type I errors for different sample sizes and test functions. These results show that the type I errors are handled satisfactorily for all the function types and sample sizes.

Power and prior sensitivity analysis

To evaluate the power and performance of FFW on DNAm data, we used the same dataset as in Lee and Morris [5]. This dataset is a combination of 26 methylation profiles on chromosome 3, containing a total of 75,069 probes. Every patient's methylation profile was measured twice, once in cancer cells and once in control cells. The phenotype (Y) is thus a binary indicator corresponding to a cancer ($Y = 1$) or control ($Y = 0$) cell.

The simulations are designed as follows. The true mean methylation level is kept identical for all the probes except the 1901 loci that were previously found to be differentially methylated in Irizarry et al. [18]. For these 1901 probes, the mean methylation levels were made to be different between cases and controls according to the difference reported by Irizarry and colleagues [18].

The above simulations are designed to ensure that the two groups have the same DNAm profile for all CpGs except the 1901 loci reported to be differentially methylated in Irizarry et al. [18]. For further information regarding the simulated data, readers are referred to the Supporting Information section in Lee and Morris [5].

The dataset itself is available at http://odin.mdacc.tmc.edu/~jmorris/simulated_data.Rdata.

Pre-processing As CpG sites are not evenly spaced in the genome, the wavelet transform is well-suited for modeling such sites as a function, provided there is a sufficiently large number of measurements. We pre-processed the DNAm data by dividing the genome into smaller regions containing at least 10 CpGs, with any two adjacent CpGs separated by a maximum distance of 500 base pairs. This criterion is similar to the one used by Jenkinson et al. [19], where the authors studied regions of 3 Kb containing at least 10 CpGs.

The above pre-processing step resulted in a total of 1213 regions, containing 1875 of the 1901 CpGs in Irizarry et al. [18] that were scattered across 89 of the 1213 defined regions. For each region, we investigated whether the CpG patterns varied between case ($n = 13$) and control ($n = 13$) cells. As each region contains at least 10 CpGs, we used a depth of analysis of 3. We then ran FFW and WaveQTL for different values of the standard deviation of the prior on the previously defined regions. Finally, for each method and standard deviation of the prior, we computed the p value for each region and the corresponding false discovery rate (FDR) using the Benjamini–Hochberg procedure [20]. Figure 3 shows the consistency of FFW according to different standard deviations of the prior. This figure also shows that FFW and WaveQTL have similar power.

To evaluate the type I error of FFW for various standard deviations of the prior, we used the above dataset of cancer and control cells to generate the test statistics under

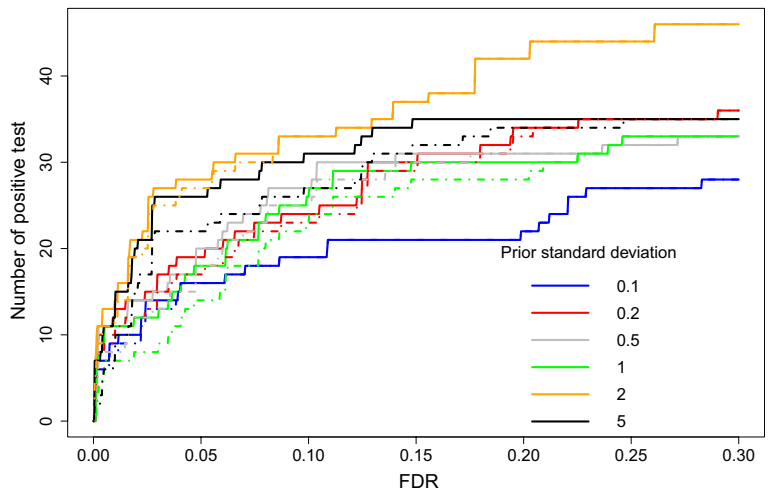

Fig. 3 Assessing the sensitivity of the standard deviations of the prior. Each curve represents the number of DNAm regions detected at a given FDR and according to different standard deviations of the prior. The solid curves are the output of FFW; the dashed curves are the output of WaveQTL

Table 1 Estimated type I error for different sample sizes and test functions

Test function	n	α **level**			
		0.5000	**0.0100**	**0.0010**	**0.0001**
Block	100	0.49980	0.00992	0.00111	0.00017
Block	500	0.50048	0.00978	0.00105	0.00009
Block	1000	0.49950	0.01028	0.00090	0.00014
Bump	100	0.50040	0.01026	0.00111	0.00004
Bump	500	0.50182	0.01040	0.00080	0.00009
Bump	1000	0.50050	0.00961	0.00098	0.00008
HeaviSine	100	0.50007	0.00982	0.00085	0.00009
HeaviSine	500	0.49956	0.00975	0.00091	0.00007
HeaviSine	1000	0.49965	0.01026	0.00103	0.00009
Doppler	100	0.50054	0.01014	0.00097	0.00009
Doppler	500	0.50089	0.01007	0.00100	0.00008
Doppler	1000	0.49916	0.01037	0.00095	0.00013

Table 2 Estimated type I error for different standard deviations of the prior

Prior σ_b	α **level**			
	0.0500	**0.0100**	**0.0010**	**0.0001**
0.1	0.0416	0.0183	0.0016	0.0008
0.2	0.0451	0.0145	0.0019	0.0012
0.5	0.0451	0.0236	0.0020	0.0010
1	0.0478	0.0119	0.0019	0.0009
2	0.0500	0.0110	0.0017	0.0009
5	0.0522	0.0210	0.0017	0.0008

the null. We performed 50 screenings of the dataset by permuting the phenotype in each screening. This corresponds to 56,500 observations of the test statistics under the null for different standard deviations of the prior. Table 2 summarizes the estimated type I errors for different standard deviations of the prior. The calibration is slightly worse than the one displayed in Table 1, which might be because the approximation used here is only valid asymptotically (see Eq. 11). Moreover, only 26 individuals were available for analysis in the case-control dataset, and we only considered independent samples in our modeling. Even though the dataset contained repeated measurements, the estimations were similar across the different priors. As was the case with statistical power, Fig. 3 also shows that FFW and WaveQTL have a similar performance.

We assessed the power of FFW according to the number of differentially methylated CpGs per region (Table 3). We computed the average proportion of truly associated regions across different standard deviations of the prior for each FDR level and the number of differentially methylated CpGs per region. FFW had higher power for regions containing a large number of differentially methylated CpGs (Table 3). We also computed the power according to the number of differentially methylated CpGs per region using WaveQTL. We found the same power estimates as those shown in Table 3. This is unexpected, considering the relatively small number of truly associated regions ($n = 89$). Figure 3 also shows slight differences between FFW and WaveQTL in relation to FDR. Still, these discrepancies are negligible and indicate that the two approaches have similar power (Additional file 1). Figure 4 shows matching ROC curves for FFW and WaveQTL, which indicate that the two methods have the same power. We suspect that, since WaveQTL estimates the p value using an early-stopping Monte Carlo approximation based on 10,000 permutations, the estimated FDR might be slightly more conservative than the one obtained based on simulations. However, when we rank the regions by FDR, we obtained the same ranking for WaveQTL as for FFW. As ROC curves are invariant if the ranking of the regions does not change, we obtain the same ROC curves as a result.

To compare FFW with another wavelet-based method, we repeated the same analyses using the "Wavelet-based Functional Mixed Models" (WFMM) method by Morris and Carrol [7] on the same dataset as above. WFMM can be used to detect DMRs [5], and, more generally, to detect signals via wavelet regression. The authors used an empirical Bayes approach to perform a regularization of the estimated effects. Their model can

Table 3 Estimated power according to the number of differentially methylated CpGs per region

FDR	Number of CpGs			
	1–10	**11–20**	**21–30**	**≥ 30**
0.01	0.000	0.101	0.094	0.385
0.05	0.117	0.182	0.254	0.58
0.10	0.150	0.256	0.348	0.564
0.15	0.233	0.291	0.370	0.564
0.20	0.300	0.302	0.370	0.577

thus take into account a larger range of correlations between the observed DNAm profiles in each individual. WFMM is thus able to handle repeated measures of DNAm.

WFMM processed all the 75,069 CpG sites in one go and computed the posterior probability for each of the CpGs being above a set threshold (here 0.05) for being associated with cancer. We ran WFMM by specifying the correlation structure between the observations. Following the approach of Lee and Morris [5], we transformed the posterior probabilities of the CpGs into Bayesian FDR [5]. To compare the performance of WFMM against that of FFW, we first need to provide a regional significance criterion for WFMM. We used the minimum Bayesian FDR value for all the CpGs within a region of interest to assign a regional significance criterion. After running WFMM on the entire dataset, we used the minimum Bayesian FDR value for each of these regions to assess significance.

The results showed that WFMM had higher power than FFW, detecting all the 89 regions with an FDR below 0.01. This difference in power might be due to the refined modeling proposed by Morris and Carrol [7], which takes advantage of the correlations between DNAm profiles. Notably, Lee and Morris [5] showed that taking these correlations into account resulted in a systematic gain in power. In terms of computational time, however, WFMM took more than 6 h to complete the screening of 1213 regions, whereas FFW took a minute.

Discussion

This paper reports on a computational shortcut for improving the wavelet-based approach proposed by Shim and Stephens [8]. We drew inspiration from the work of both Shim and Stephens [8] and Zhou and Guan [10] to develop a faster functional modeling that is applicable to a wider variety of functions and phenotypes. The approach of Shim and Stephens [8] was designed to identify dsQTL. Here, we show that wavelet-based approaches can also be used to detect differentially methylated regions (DMRs). Both WaveQTL by Shim and Stephens [8] and FFW offer a more flexible approach to modeling functions than conventional single-point testing. By keeping the design matrix constant across the screened regions and using simulations instead of permutations, we show that FFW is faster than WaveQTL. In addition, FFW controls the type I error satisfactorily for large sample sizes.

Reverse regression is a very useful tool for reducing the overall computational burden. However, the downside of reverse regression is that the coefficients from the analysis may become less interpretable. If the objective of a study is to estimate the effect of a particular wavelet in a given DNAm region, then one needs to rerun the procedure using individual wavelet coefficients as exposures. Therefore, FFW might function better as an initial screening tool to gain important biological insights from the DNAm data. For other types of data, such as those from biomechanical research, the wavelet coefficients are more directly interpretable. For example, if a researcher is interested in studying the effect of a particular treatment on motor function, e.g., leg function in the strength-dexterity test [21], FFW would lend itself easily to such an analysis.

An additional methodological constraint is the need to assign a given value for J in the applications of FFW. We chose a cutoff of $J = 3$ because of the requirement for the screened regions to contain least 10 CpGs, with any two adjacent CpGs separated

by a maximum distance of 500 base pairs. In general, we advise choosing as large of a J as possible, while restricting $2^J < \kappa$, where κ is an integer larger than 4. This can be written as $J = max_j\{2^j < \kappa\}$ when analyzing regions containing at least κ CpGs. In our current application, this corresponded to $J = 3$. Nevertheless, it is possible to choose a different J for each region. Given the test statistic depends on J, choosing a different J for each genetic region would require the test statistic to be simulated for different values of J. As shown in Fig. 1, one can quickly simulate the test statistic. Therefore, simulating different values of J is likely to have a negligible impact on the overall run time.

FFW is well-powered to detect DMRs containing more than 10 CpGs, even when the CpGs only have small effects. However, FFW has lesser power for detecting DMRs containing only a few CpGs (≤ 10). Although it is less powerful than WFMM [5], FFW has the advantage of being significantly faster in terms of computational time. WFMM took more than 6 h to process one chromosome for 26 individuals, whereas FFW took a minute. We thus expect WFMM to become exceedingly slow if there is a need to scale up the analysis to include hundreds of individuals and data from denser DNAm platforms, such as the Illumina 850K, or those from whole-genome bisulfite sequencing [22]. Therefore, FFW is a useful complementary tool for the rapid scanning of EWAS datasets to detect DMRs that can subsequently be used in downstream fine-mapping efforts. An attractive application of FFW would be to re-analyze DNAm data from previously published EWASes that are publicly available through, e.g., the Gene Expression Omnibus (GEO) database [23].

In future developments, we plan to extend FFW to also include phenotypes on non-ordered scales, e.g., blood types and psychiatric phenotypes. Such phenotypes are routinely treated in a case-control fashion and analyzed using multinomial regression owing to the prohibitively large computational burden. However, by exploiting reverse regression, as we do here, the phenotypes can be re-coded and readily included in the predictor matrix. Reverse regression also enables FFW to easily adapt to the setting of a phenome-wide association study (PheWAS), in which multiple phenotypes are interrogated simultaneously ([24–26]). As highlighted by our analyses, this development is further simplified by the results of Zhou and Guan [10], showing that the parameter of the Bayes factors law depends primarily on the singular values of the regression matrix and the number of parameters tested. As the regression matrix remains constant across all loci, locations, and scales, these parameters only need to be computed once, thus enabling a fast computation of p values. This makes FFW a highly versatile method for analyzing phenotypes that do not lend themselves easily to either single-point or bump-hunting methods.

FFW is distributed as an ***R*** package. The package contains the analysis code and a data visualization tool to enable a more detailed inspection of the detected regions. The full ***R*** package is freely available on GitHub (https://github.com/william-denault/ffw), and a comprehensive example run of the package is provided in the help function *ffw*.

Appendix

Computational cost

Here, we discuss the improvement in computational cost using FFW. Let R be the number of regions to screen for associations (i.e., the number of tests to be performed). When considering a Bonferroni correction for R tests, the number of permutations or simulations needed to obtain sufficiently precise p values for a multiple testing correction of $\frac{0.05}{R}$ is of the order $O(R^2)$ (see Knijnenburg et al. [27]). The computational cost of performing a linear regression is $O(np^2 + p^3)$, where n is the number of observations and p is the number of variables. For every screened region, we perform 2^J linear regressions. The complexity of the wavelet transform is equal to the number of observed variables, which is 2^J in our case. It follows that the combined computational cost using WaveQTL for a full screening of R regions and the permutation procedure is:

$$O\left(2^J\left(np^2 + p^3 + e\right)\left(1 + R^2\right)R + 2^J R\right) \tag{14}$$

where e is the average number of iterations of the EM algorithm. Using early-stopping methods described by Gandy [28] and Besag and Clifford [29] for deriving Monte Carlo p values, the computational cost can be further reduced to:

$$O\left(2^J\left(np^2 + p^3 + e\right)(1 + S)R + 2^J R\right) \tag{15}$$

where S is the expected number of steps of the early-stopping Monte Carlo p values method, with $S \leq R^2$. The first term in Eq. (15) is the number of regressions to be performed, including the ones for the permutation as well as for the cost of the EM algorithm. The second term is the overall cost of the wavelet transform.

Early-stopping methods, such as the methods of Gandy [28] or Besag and Clifford [29], are algorithms that estimate p values in a sequential fashion. As in most applications, the interesting p values are the small p values. Early-stopping methods aim at avoiding using a lot of computational power on large p values. Therefore, when it is likely that the test p value is large, such algorithms stop and move on to the next p value to be estimated. Early-stopping methods thus aim at focusing only on the interesting tests and allocating most of the computational power to those. The *simctest R* package contains most of recent early-stopping algorithms, and include the method of Gandy [28].

Using FFW, and neglecting the cost of simulations, we obtain a computational cost of:

$$O\left(2^J\left(np^2 + p^3\right)R + 2^J R + e\left(R + R^2\right)\right) \tag{16}$$

where the first term is the computational cost of all the regressions performed, the second term is the overall cost of the wavelet transform, and the third term is the number of iterations performed by the EM algorithm. The R term is for running the EM algorithm for each region, and the R^2 term is for running the EM algorithm for each simulation used to assess the significance of each region. Using early-stopping rules for Monte Carlo p values, this can be further reduced to:

$$O\left(2^J\left(np^2 + p^3\right)R + 2^J R + e\left(R + S'\right)\right) \tag{17}$$

where S' is the number of iterations needed to assess the significance of the smallest p value using the early-stopping rules for Monte Carlo p values ($S' \leq R^2$). In essence, this procedure reduces the degree of the polynomial cost of the region by one. In addition, the term $\left(np^2 + p^3\right)$, which carries part of the computational burden, is now only associated with a first-order polynomial in R and no longer associated with a third and second-degree polynomial. In a scenario in which all the regions have 2^J variables, the first term would thus represent the overall cost corresponding to the single-point testing strategy.

Abbreviations

DMR: Differentially methylated region; DNAm: DNA methylation; dsQTL: DNase I hypersensitivity quantitative trait locus; EWAS: Epigenome-wide association study; FDR: False discovery rate; GEO: Gene Expression Omnibus; GWAS: Genome-wide association study; MCMC: Markov Chain Monte Carlo; NIG: Normal-inverse gamma; PheWAS: Phenome-wide association study; QTL: Quantitative trait locus.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-021-03979-y.

Additional file 1: Figure S4. ROC curves at given standard deviations of the prior. The thin solid curves are the output of FFW; the thick dashed curves are the output of WaveQTL. The ROC curves match for standard deviations of the prior of 0.5 and 1.

Acknowledgements

We thank Dr. Håkon K. Gjessing for his valuable comments on earlier drafts of the manuscript. We also thank Drs. Lee and Morris [5] for making their DNAm data freely available to the wider research community.

Authors' contributions

WD and AJ conceptualized the study. WD proposed the use of the approach by Zhou and Guan [10] to enhance the computation of Shim and Stephens [8] and drafted the first version of the manuscript. WD also wrote the R package and performed the simulations. AJ provided overall scientific input as well as editorial assistance in drafting the manuscript to its current form. Both authors read and approved the final manuscript.

Funding

This project was funded in part by a grant from the Research Council of Norway (RCN) (Grant 249779). Additional funding was provided by RCN through its Centres of Excellence funding scheme (Grant 262700). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Availability of data and materials

The full R package for FFW is freely available on GitHub (https://github.com/william-denault/ffw), and a comprehensive example run of the package is provided in the help function ***ffw***. The data supporting the findings of this study are openly available at http://odin.mdacc.tmc.edu/~jmorris/simulated_data.Rdata.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Genetics and Bioinformatics, Norwegian Institute of Public Health, Oslo, Norway. [2]Centre for Fertility and Health, Norwegian Institute of Public Health, Oslo, Norway. [3]Department of Global Public Health and Primary Care, University of Bergen, Bergen, Norway.

Received: 29 September 2020 Accepted: 27 January 2021
Published online: 10 February 2021

References

1. Cahouët V, Luc M, David A. Static optimal estimation of joint accelerations for inverse dynamics problem solution. J Biomech. 2002;35(11):1507–13. https://doi.org/10.1016/S0021-9290(02)00176-8.

2. Degner JF, Pai AA, Pique-Regi R, Veyrieras J-B, Gaffney DJ, Pickrell JK, De Leon S, Michelini K, Lewellen N, Crawford GE, Stephens M, Gilad Y, Pritchard JK. DNase I sensitivity QTLs are a major determinant of human expression variation. Nature. 2012;482(7385):390–4. https://doi.org/10.1038/nature10808.
3. Vsevolozhskaya OA, Zaykin DV, Greenwood MC, Wei C, Lu Q. Functional analysis of variance for association studies. PLoS ONE. 2014;9(9):105074. https://doi.org/10.1371/journal.pone.0105074.
4. Jaffe AE, Murakami P, Lee H, Leek JT, Fallin MD, Feinberg AP, Irizarry RA. Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. Int J Epidemiol. 2012;41(1):200–9. https://doi.org/10.1093/ije/dyr238.
5. Lee W, Morris JS. Identification of differentially methylated loci using wavelet-based functional mixed models. Bioinformatics. 2016;32(5):664–72. https://doi.org/10.1093/bioinformatics/btv659.
6. Friedman JH, Fisher NI. Bump hunting in high-dimensional data. Stat Comput. 1999;9(2):123–43. https://doi.org/10.1023/A:1008894516817.
7. Morris JS, Carroll RJ. Wavelet-based functional mixed models. J R Stat Soc Ser B (Stat Methodol). 2006;68(2):179–99. https://doi.org/10.1111/j.1467-9868.2006.00539.x.
8. Shim H, Stephens M. Wavelet-based genetic association analysis of functional phenotypes arising from high-throughput sequencing assays. Ann Appl Stat. 2015;9(2):665–86. https://doi.org/10.1214/14-AOAS776.
9. Ma L, Soriano J. Efficient functional ANOVA through wavelet-domain Markov groves. J Am Stat Assoc. 2018;113(522):802–18. https://doi.org/10.1080/01621459.2017.1286241.
10. Zhou Q, Guan Y. On the null distribution of Bayes factors in linear regression. J Am Stat Assoc. 2017;113(523):1362–71. https://doi.org/10.1080/01621459.2017.1328361.
11. Fuller W. Measurement error models. 1st ed. Hoboken: Wiley; 1987. https://doi.org/10.1002/9780470316665.
12. Nason G. Wavelet methods in statistics with R. Use R! New York: Springer; 2008. https://doi.org/10.1007/978-0-387-75961-6.
13. Kovac A, Silverman BW. Extending the scope of wavelet regression methods by coefficient-dependent thresholding. J Am Stat Assoc. 2000;95(449):172–83. https://doi.org/10.2307/2669536.
14. Nason GP, Sachs RV, Kroisandt G. Wavelet processes and adaptive estimation of the evolutionary wavelet spectrum. J R Stat Soc Ser B (Stat Methodol). 2000;62(2):271–92. https://doi.org/10.1111/1467-9868.00231.
15. Mallat S.G. A wavelet tour of signal processing: the sparse way. 3rd ed. Amsterdam: Elsevier/Academic Press; 2009. https://doi.org/10.1016/B978-0-12-374370-1.X0001-8.
16. Jentsch C, Kirch C. How much information does dependence between wavelet coefficients contain? J Am Stat Assoc. 2016;111(515):1330–45. https://doi.org/10.1080/01621459.2015.1093945.
17. Donoho DL, Johnstone IM. Ideal spatial adaptation by wavelet shrinkage. Biometrika. 1994;81(3):425–55. https://doi.org/10.1093/biomet/81.3.425.
18. Irizarry RA, Ladd-Acosta C, Wen B, Wu Z, Montano C, Onyango P, Cui H, Gabo K, Rongione M, Webster M, Ji H, Potash J, Sabunciyan S, Feinberg AP. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nat Genet. 2009;41(2):178–86. https://doi.org/10.1038/ng.298.
19. Jenkinson G, Pujadas E, Goutsias J, Feinberg AP. Potential energy landscapes identify the information-theoretic nature of the epigenome. Nat Genet. 2017;49(5):719–29. https://doi.org/10.1038/ng.3811.
20. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Stat Soc Ser B (Methodol). 1995;57(1):289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.
21. Lyle MA, Valero-Cuevas FJ, Gregor RJ, Powers CM. Control of dynamic foot-ground interactions in male and female soccer athletes: Females exhibit reduced dexterity and higher limb stiffness during landing. J Biomech. 2014;47(2):512–7. https://doi.org/10.1016/j.jbiomech.2013.10.038.
22. Suzuki M, Liao W, Wos F, Johnston AD, DeGrazia J, Ishii J, Bloom T, Zody MC, Germer S, Greally JM. Whole-genome bisulfite sequencing with improved accuracy and cost. Genome Res. 2018;28(9):1364–71. https://doi.org/10.1101/gr.232587.117.
23. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A. NCBI GEO: archive for functional genomics data sets-update. Nucleic Acids Res. 2013;41(D1):991–5. https://doi.org/10.1093/nar/gks1193.
24. Liu Z, Lin X. Multiple phenotype association tests using summary statistics in genome-wide association studies. Biometrics. 2018;74(1):165–75. https://doi.org/10.1111/biom.12735.
25. Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry K, Wang D, Masys DR, Roden DM, Crawford DC. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics. 2010;26(9):1205–10. https://doi.org/10.1093/bioinformatics/btq126.
26. Verma A, Lucas A, Verma SS, Zhang Y, Josyula N, Khan A, Hartzel DN, Lavage DR, Leader J, Ritchie MD, Pendergrass SA. PheWAS and beyond: the landscape of associations with medical diagnoses and clinical measures across 38,662 individuals from Geisinger. Am J Hum Genet. 2018;102(4):592–608. https://doi.org/10.1016/j.ajhg.2018.02.017.
27. Knijnenburg TA, Wessels LFA, Reinders MJT, Shmulevich I. Fewer permutations, more accurate P-values. Bioinformatics. 2009;25(12):161–8. https://doi.org/10.1093/bioinformatics/btp211.
28. Gandy A. Sequential implementation of Monte Carlo tests with uniformly bounded resampling risk. J Am Stat Assoc. 2009;104(488):1504–11. https://doi.org/10.1198/jasa.2009.tm08368.
29. Besag J, Clifford P. Sequential Monte Carlo p-values. Biometrika. 1991;78(2):301–4. https://doi.org/10.2307/2337256.

Publisher's Note