Pacific Symposium on Biocomputing 12:201-215(2008)



# MULTI-SCALE CORRELATIONS IN CONTINUOUS GENOMIC DATA

R. E. THURMAN, W. S. NOBLE, AND J. A. STAMATOYANNOPOULOS

*Department of Genome Sciences*
*University of Washington*
*1705 NE Pacific St*
*Seattle, WA 98195–5065*

Functional genomic quantities such as histone modifications, chromatin accessibility, and evolutionary constraint can now be measured in a nearly continuous fashion across the genome. The genome is highly heterogeneous, and the relationships between different functional annotations may be fluid. Here we present an approach for visualizing, quantifying, and determining the statistical significance of local and regional correlations between high-density continuous genomic datasets. We use wavelets to generate a multi-scale view of each component data set and calculate correlations between data types as a function of genome position over a continuous range of scales in sliding window fashion. We determine the statistical significance of correlations using a non-parametric sampling approach. We apply the wavelet correlation method to histone modification and chromatin accessibility (DNaseI sensitivity) data from the NHGRI ENCODE project. We show that DNaseI sensitivity is broadly correlated (though to differing degrees) with a number of different activating histone modifications. We examine the continuous relationship between the repressive histone modification H3K27me3 and the activating mark H3K4me2, and find these modifications to display significant duality, with both significant positively and negatively correlated genomic territories. While the former appear to recapitulate in definitive cells the so-called "bi-valent" pattern originally proposed as a signature of pluripotency, the presence of negatively correlated regions suggests that the regulatory events that underlie the observed modification patterns are complex and highly regionalized in the genome.

## 1. Introduction

Rapid progress in the development and application of high-density functional genomic assays has spawned a deluge of new data types. This in turn has created a significant need for computational tools to assess quantitatively the relationship between different data types as a function of genomic position, in a manner that can be related to existing genomic annotations such as genes and transcripts. Data types now available through

Pacific Symposium on Biocomputing 12:201-215(2008)



large-scale efforts such as the NHGRI ENCODE project include various histone modifications, chromatin accessibility, DNA replication timing, bulk transcriptional/RNA output, and evolutionary conservation. Since none of these data have been available in a continuous fashion across diverse genomic regions, their interrelationships are largely unknown. For instance, how does transcriptional activity relate to replication timing? Is the relationship constant across the genome, or is it regionalized? How do histone modifications relate to chromatin accessibility and transcription, particularly in intergenic regions? How does this relationship vary over different parts of a gene, or between gene-rich and gene-poor regions? The human genome is functionally heterogenous, and the pending availability of genome-wide data sets render these questions highly relevant to our understanding of the functional architecture of the genome. The increasing scope and resolution of high throughput genomic assays encourages a multi-scale view of the genome, where some processes vary rapidly over tens or hundreds of bases, and others vary slowly over tens or hundreds of kilobases. We therefore desire to view functional genomic activity over a wide range of scales that may evince both nucleotide- and domain-level phenomena [15].

Here we present a method based on wavelet analysis for simultaneously computing and displaying correlations between different continuous genomic data types at multiple scales. Wavelets provide a mathematical framework for analyzing time series-like data at multiple scales. In the parlance of signal-processing, wavelets are a fundamental tool for *time-frequency* analysis, which in the context of genomic data means that they can describe features in data that are both scale-specific and position-specific (see Methods, below).

Briefly, our method consists first of computing the continuous wavelet transform over a range of scales for each of a pair of datasets to be compared. This gives a multi-scale representation of each dataset, as well as normalizing each pair to a common set of scales. We then correlate the wavelet-transformed results in sliding window fashion on a scale-by-scale basis. The resulting correlation patterns can be visualized in heatmap form, or in aggregate using histograms. We assess the statistical significance of these patterns using non-parametric methods including the Kolmogorov-Smirnov test and, primarily, sampling techniques.

The results and approach presented here expand on those developed in the pilot phase of the ENCODE project [8, 9], whose mission is to identify all functional elements in the human genome. A distinguishing feature of the ENCODE project is its charge to encompass a large number of di-

Pacific Symposium on Biocomputing 12:201-215(2008)



verse data types collected using high-throughput techniques (tiling DNA microarrays, high-throughput real-time PCR, and, more recently, ultra-high-throughput sequencing) that collectively expose the functional activity of the human genome in vivo.

The techniques presented here and previously [9] form a key part of the effort required to integrate these diverse functional quantities. In the present work we use improved wavelet techniques and a more rigorous statistical framework to analyze in greater detail the relationships between histone modifications and chromatin accessibility, and between different histone modification classes. The latter have attracted significant attention recently with the observation that certain modifications occurring in combination (H3k27me3 and H3k4me2/3) may have a special functional significance for cellular state.

Chromatin accessibility is a first-order indicator of chromatin structure, and it has long been measured by quantifying sequence-specific or regional DNaseI sensitivity. Epigenetic factors such as histone modifications are thought to play key roles in a number of biological processes, including initiation and propagation of transcription, and higher-order chromatin organization. Nevertheless, the interrelationships between the different histone modifications and chromatin accessibility have not been systematically studied prior to the availability of the ENCODE data.

## 2. Related work

Wavelets have been used in a number of bioinformatics applications to detect and analyze patterns in sequence data [11]; to de-noise microarray data [11]; and to elucidate large-scale trends in functional genomic data [15]. Wavelets have also been used to uncover sequence and gene-related correlations between prokaryotic species [1]. In these studies, correlations were measured by identifying regions (in position-scale space) of significant wavelet coefficients that were shared between datasets. By contrast, the approach presented here measures the correlation between wavelet coefficients directly, in sliding window fashion across a chromosome. In [5] the authors compare observed and randomized histograms of local correlation coefficients to relate the divergence in non-coding non-repetive DNA with the amount of repetitve DNA.

Pacific Symposium on Biocomputing 12:201-215(2008)



## 3. Results

We present two analyses to illustrate our methods. Where noted, Supplementary Information is available at `http://noble.gs.washington.edu/proj/wavecor`.

### 3.1. *Correlated data: DNaseI vs. H3K4me2*

The pilot phase of ENCODE [9] focused on a representative cross-section of 1% of the genome (approximately 30Mb), divided into 44 regions. One of the main conclusions drawn from the pilot analysis is that chromatin accessibility, as measured by DNaseI sensitivity, is very broadly correlated with activating histone modifications including bulk acetylation of histones H3 and H4 (H3ac, H4ac), and mono-, di-, and tri-methylation of H3 lysine 4 (H3K4me1/2/3). Figure 1 shows a heatmap depicting the correlation betweeen DNaseI and H3K4me2, which were jointly measured in the GM06990 lymphoblastoid cell line. The following steps outline the process for generating, interpreting, and assessing the statistical significance of this correlation map. The chromatin accessibility data were generated using the DNase-array method [14], and the histone modification data derived from the Sanger Institute [10] ENCODE studies.

#### 3.1.1. *Wavelet coefficients*

The continuous wavelet transform (CWT; see Methods) coefficient for a given dataset can be computed at any position and any scale greater than the resolution of the input data. The CWT encapsulates how much the data are changing at that scale and position. For correlation analysis, we compute the CWT coefficients at a range of scales for each dataset across the regions of interest. This procedure results in a matrix of CWT coefficients for each dataset, with the x-axis representing genomic position and the y-axis representing wavelet scale. Figure 2 shows heatmap representations (or *scalograms*) of CWT matrices for DNaseI sensitivity and H3K4me2 at scales ranging from 2kb to 32kb. The wavelet family used here is an improved version of that used in [9] in its ability to capture more accurately negative correlations between the data types (see Methods).

#### 3.1.2. *Correlation heatmaps*

The scalograms in Figure 2 show marked similarity in both position and scale; it is these similarities that we aimed to quantify. We computed the

Pacific Symposium on Biocomputing 12:201-215(2008)



Pearson correlation of the CWT coefficients at each scale, in a sliding window fashion across the genome. Figure 1 shows a heatmap representation of this matrix for DNaseI and H3K4me2 using a sliding window with width at any given scale equal to 2.5 times the scale (e.g., 25kb window at 10kb scale). The high percentage area of red in this figure is qualitative evidence of a high degree of positive correlation at multiple scales.

The width of the sliding window is arbitrary, and can be tailored to fit, for example, prior knowledge of the scale of effect of a biological phenomenon (e.g., the size of a nucleosome, the size of the average gene, etc.). However, wider windows may defeat the purpose of isolating local correlations, while shorter windows push correlation values towards the extremes of $\pm 1$. This latter effect occurs also for a fixed window width as the scale increases. The scale-adaptive width used here makes correlations comparable across scales, in contrast to the fixed window size technique used previously [9].

#### 3.1.3. *Statistical significance*

We next addressed the statistical significance of the preliminary conclusions brought forth by visual inspection of Figure 1. Specifically, how significant is the observed global positive correlation over random expectation? Is this correlation profile more or less extreme when we replace H3K4me2 by another histone modification? Do the results change if we consider all 44 ENCODE regions? To address these questions we applied non-parametric methods: the Kolmogorov-Smirnov (KS) test [6] for assessing the differences between distributions, and iterative random sampling to form empirical null distributions.

**Statistical significance via KS test.** Figure 3, left, shows the smoothed histogram of ENCODE-wide sliding window correlation values between DNaseI and each of the five histone modifications at the 16kb scale. The high degree of positive correlation displayed in Figure 1 is reflected in Figure 3, where the distribution for all marks is highly skewed toward $+1$. The distributions appear, moreover, to be ordered with respect to the degree of positive skew, with H3K4me2 most correlated, followed by H3K4me1, H3K4me3, H3ac, and H4ac in that order. Application of the one-sided KS test showed that the ordering is significant for all five marks ($p < 10^{-29}$), except for the relationship between H3K4me1 and H3K4me3, which is ambiguous. Results from [9] showed H3K4me2, H3K4me3 and H3ac most

Pacific Symposium on Biocomputing 12:201-215(2008)



correlated with DNaseI, and that group being significantly more correlated than H3K4me1 and H4ac, but with no other significant ordering of the marks within those groups possible. Taken together, these results provide strong evidence for high though graded correlation between DNaseI sensitivity and the sampled range of histone modifications.

**Statistical significance via sampling.** The techniques in the previous section may be used to compare different sets of observed correlations. Next we addressed the question of comparing observed correlations to random expectation. Here, "random expectation" means relative to a null distribution formed by considering two random signals that are, in a critical sense, similar to the two observed datasets.

For the null model, rather than fitting a parametric model of the signals, which involves assumptions or simplifications of the data that may be incorrect, we pursued a non-parametric sampling approach. All available data for a given time series were concatenated into a single master series which served as a pool from which regions of fixed size were sampled (with size depending on the question being asked). Each point in the null distribution was derived by computing the correlation between regions sampled from independent positions in the two master series. This technique maintains the internal structure of the original time series while breaking any correlation between them. We obtained an empirical p-value by counting how many points in the null distribution met or exceeded the properties of the observed feature.

Figure 3, right, shows the distribution of all observed DNaseI/H3K4me2 sliding window correlation values in the 500kb region ENr132 at the 16kb scale. We found that 52% of the correlation coefficients at this scale exceed 0.7. To calculate significance, we randomly sampled blocks of size 500kb from each master series at the 16kb scale. Out of 5,000 sampled correlation profiles, only two had at least 52% of their values over 0.7, yielding a $p$-value of 0.0004 for this region. Figure 4, center, shows a plot of additional sampled correlation profiles.

There are 31 ENCODE regions of size exactly 500kb. We repeated the above analysis for each of these regions and obtained uncorrected empirical p-values ranging from 0.0000 to 0.3948 (see Supplementary Information). The variability in these results suggests region-specific differences affecting the correlations between DNaseI and histone modifications. For example, Figure 5 shows that regions of high gene density tend to have higher correlation values.

Pacific Symposium on Biocomputing 12:201-215(2008)



When ENCODE-wide data were considered (Figure 3, left), we observed that 56% of all 16kb DNaseI/H3K4me2 correlation values exceeded 0.7. To test the significance of findings in the wider data set, we computed a null distribution in which each point derived from sampling (for each dataset), 44 ENCODE region-sized pieces (with replacement) from the master series. Of 1,000 samples so obtained, none achieved a degree of positive correlation comparable to the observed (empirical p-value of $< 0.001$). This result supports the intuition that chance high correlation (pseudocorrelation) in random data is significantly harder to sustain over longer regions than shorter regions. Figure 4 shows the density of several ENCODE-wide sampled correlation profiles versus several 500kb sampled profiles, with the latter evincing far more variability. Correspondingly, Figure 4, right, shows that the size of the tails of each sampled correlation profile has a much wider distribution for the 500kb samples than for the ENCODE-wide samples.

### 3.2. *Uncorrelated data: H3K4me2 vs. H3K27me3*

As a further illustration of the utility of this approach, and to introduce additional methods, we next examined the relationship between two histone modifications, H3K4me2 and H3K27me3. The former is classically associated with transcriptional activation, while the latter is held to signify transcriptional or even regional chromatin repression. Due to the dual nature of these modifications we expected their profiles in lymphoblastoid cells to be largely uncorrelated, or perhaps even anticorrelated. Indeed, Figure 6, top, covering the alpha globin cluser (Chr16) shows clearly co-located peaks near position Chr16:150,000. This location contains a block of high positive correlation across multiple scales, while a number of flanking peaks for one mark or the other show no correlation or anticorrelation. Numerous examples of analogous co-located peaks occur throughout the ENCODE regions, as do examples of slightly offset peaks (see Supplementary Information for ENCODE-wide plots). The co-localization of H3K4me2/3 with H3k27me3 was first described in mouse embryonic stem cells (where it is prominent over the promoters of certain developmentally coordinated genes), and was originally thought to be a marker of pluripotency [4, 2]. However, more recent work has called this conclusion into question [3].

Figure 6 (bottom, solid line) shows the distribution of ENCODE-wide sliding window correlation values for these two marks at the 16kb scale. The plot reveals a fairly balanced distribution of positive and negative correlation values. Viewed independently, it is not clear whether this pattern

is indicative of random non-correlation or if it reflects the aggregate of non-random patterns of real positive and negative correlations, including the bivalent-like domains suggested at the top of Figure 6.

To address this issue, we took two approaches. First, we performed sampling experiments to ascertain null behavior. We calculated correlation profiles from 1,000 ENCODE-wide random samples of the two sets of 16kb wavelet coefficients. Several correlation profiles are plotted as dashed lines in Figure 6. While the sampled distributions are also largely balanced between positive and negative values, the observed distribution shows extension of the tails and an offsetting central depression. Indeed, we found that no sampled distribution had the same fraction of coefficients above 0.5 and below $-0.5$ as the observed, for an empirical p-value of $< 0.001$. This provides quantitative evidence for non-random positive and negative correlations.

Next, we attempted to identify regions of local agreement between the two marks. We performed a 2-state HMM segmentation of each dataset, partitioning the ENCODE regions into "high" and "low" states based separately on wavelet smoothed versions of each mark at scales ranging from 4kb to 128kb (see [7] for methods). We then formed the intersection of the high states for both marks at each scale.

For data smoothed at the 4kb scale the intersection of the high states comprised approximately 3.4Mb ($> 10\%$) of ENCODE, which overlapped 70 annotated genes. GO analysis revealed 19 categories over-represented at p-values less than 0.01, including six transcription-related terms, terms for the regulation of cellular, physiological and biological processes, for phosphatase and enzyme activity, and for development (see Supplementary Information for full results). These categories accord with prior observations that a significant fraction of bivalent domains occur at genes encoding transcriptional regulatory factors or at the 3' ends of developmental genes [4]. They also found large bivalent domains in the Hox clusters.

To explore scale-specific effects, we repeated the HMM segmentations using 16kb, 64kb and 128kb scales. The intersection of the H3K27me3 and H3K4me2 high states at these scales covered 4.5Mb, 7.6Mb, and 8.0Mb, respectively. Almost without exception, the over-represented GO terms at each scale were a subset of the terms at the next smaller scale, and all terms were a subset of the 4kb terms. Five terms were over-represented at the 128kb scale, with transcription factor activity being the most significant ($p = 2.7 \times 10^{-7}$). See Supplementary Information for full results.

Pacific Symposium on Biocomputing 12:201-215(2008)



## 4. Discussion and conclusions

The influx of functional genomic data types collected using high-throughput methods has created a significant need for tools to integrate diverse signals into a meaningful picture of the functional structure of the human genome. The methods presented here are widely applicable to this problem. In the course of the ENCODE pilot data analyses we performed wavelet correlation and visualization analyses for scores of pairs of data types including, in addition to those discussed here, replication timing, evolutionary conservation measures, bulk RNA output, and nucleosome depletion assays (see results provided in Supplementary Information.

A key aspect to our approach is the systematic integration of locoregional analyses. These results can be used to confirm hypotheses concerning the relation between data types proposed on the basis of mechanistic relationships (e.g., the correlation between DNaseI sensitivity and activating histone modifications in gene-rich regions), and they may be applied in an exploratory mode, such as de novo identification of regions of common but generally unexpected high activity of activating (H3K4me2) and repressive (H3K27me3) modifications.

Future work will be required to elucidate the complex relationship between activating and repressive histone modifications. Indeed, it appears preliminarily that these broad labels are not sufficient to categorize behavior across all genomic terrain. Additionally, it is not clear to what extent the locoregional relationships between different modifications depend on the cell type being studied. While the particular genes or domains that evince a particular combination of marks may change between cell types, it will be interesting to determine whether the overall proportion of territory covered by that combination changes substantially. The pending availability of additional data from ENCODE as well as other large-scale chromatin profiling efforts [3, 12], including both additional cell types and additional modifications will provide an opportunity to address this sytematically. Ultimately, extension of the approach described here to encompass multiple diverse data types in addition to histone modifications and chromatin accessibility will likely hold the greatest promise for elucidating the functional landscape of the genome. On a practical level, many of the calculations required for the wavelet correlation approach are computationally expensive. As data types proliferate and expand to encompass the whole genome, a first priority will be to determine whether the distribution of correlation values observed are largely independent of data-type and dependent only

on scale, which would permit realization of significant efficiencies through pre-computation of case sampled distributions.

## 5. Methods

**Data normalization.** Wavelet correlation analysis requires that both datasets be defined on a common set of equally-spaced genomic coordinates. We performed gap-filling interpolation and wavelet normalization as described previously [9].

**Wavelets.** In the analysis of time series-like data, wavelet analysis can be thought of as an extension of Fourier analysis. Both techniques are used to look for periodicities or strong changes in a time series at a given period, or, in the language of wavelets, *scale*. But whereas Fourier analysis is global in nature, giving a single Fourier coefficient for each period for an entire time series, wavelet analysis is local, producing a wavelet coefficient at any point in the time series that describes the strength of the change in the time series at the given scale, at that time. We use the collection of wavelet coefficients across time (genomic position, in our case) for a fixed scale to summarize the scale-specific behavior of the time series.

Other smoothing techniques (loess, sliding window averaging) could be used to also approximate scale-specific behavior. We chose wavelets because of the availability of a computational framework (R package Rwave) for simultaneously computing wavelet coefficients across multiple scales and the established role of wavelets in time-frequency analysis.

The basis for all wavelet analysis is the continuous wavelet transform (CWT) [13]. For a given equally-spaced time series $x(t)$, the CWT wavelet coefficient $W(a, s)$ for given scale $a$ and time $s$ is given by

$$W(a,s) \equiv \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(u)\psi\left(\frac{u-s}{a}\right)\, du,$$

where $\psi(t)$ is the wavelet function of choice, satisfying the basic properties $\int_{-\infty}^{\infty} \psi(u), du = 0$ and $\int_{-\infty}^{\infty} \psi^2(u), du = 1$. We use our own implementation of the real-valued "first derivative of Guassian," or DOG wavelet. By contrast, the analysis in [9] used the complex-valued Morlet wavelet; correlations using this wavelet required taking the absolute value of the coefficients, which masked negative correlations.

## References

1. T. Allen *et al.* Long-range periodic patterns in microbial genomes indicate significant multi-scale chromosomal organization. *PLoS Computational Biology*, 2(1):13–21, January 2006.
2. V. Azuara *et al.* Chromatin signatures of pluripotent cell lines. *Nature Cell Biology*, 8:532–538, 2006.
3. A. Barski *et al.* High-resolution profiling of histone methylations in the human genome. *Cell*, 129:823–837, 2007.
4. B. E. Bernstein *et al.* A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell*, 125(2):315–326, 2006.
5. F. Chiaromonte *et al.* Association between divergence and interspersed repeats in mammalian noncoding genomic dna. *PNAS*, 98:14503–14508, 2001.
6. W. J. Conover. *Practical Nonparametric Statistics*. Wiley Series in Probability and Statistics. Wiley & Sons, 3rd edition, 1999.
7. N. Day *et al.* Unsupervised segmentation of continuous genomic data. *Bioinformatics*, 23:1424–1426, 2007.
8. ENCODE Consortium. The ENCODE (ENCyclopedia Of DNA Elements) project. *Science*, 306(5696):636–640, 2004.
9. ENCODE Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447:799–816, 2007.
10. C. Koch *et al.* The landscape of histone modifications across 1% of the human genome in five human cell lines. *Genome Research*, 17:691–707, 2007.
11. P. Liò. Wavelets in bioinformatics and compuataional biology: state of art and perspectives. *Bioinformatics*, 19(1):2–9, 2003.
12. T. S. Mikkelsen *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-commited cells. *Nature*, 448:553–560, 2007.
13. D. B. Percival and A. T. Walden. *Wavelet Methods for Time Series Analysis*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2000.
14. P. J. Sabo *et al.* Genome-scale mapping of DNaseI sensitivity *in vivo* using tiling DNA microarrays. *Nature Methods*, 3:511–518, 2006.
15. R. E. Thurman *et al.* Identification of higher-order functional domains in the human ENCODE regions. *Genome Research*, 17:917–927, 2007.

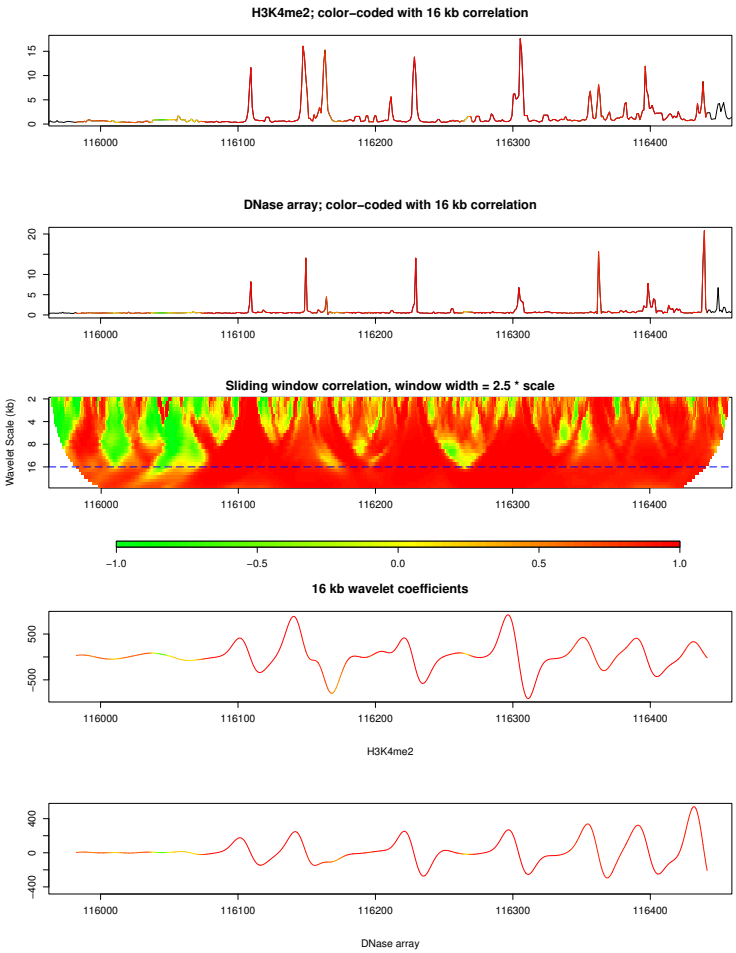


Figure 1. Correlation heatmap for H3K4me2 vs. DNaseI (GM06990) in ENCODE region ENm003. From top to bottom: raw data for H3K4me2, raw data for DNaseI, correlations heatmap, 16kb CWT coeffiecients for H3K4me2, 16kb CWT coeffiecients for DnaseI. Raw data and coefficients are colored with the correlations at the 16kb scale (dashed line).