Discovery of higher-order functional features within ENCODE regions

Robert Thurman[a], William S. Noble[b], and John A. Stamatoyannopoulos[b]

[a]Division of Medical Genetics and [b]Department of Genome Sciences, University of Washington, Seattle, WA, USA.

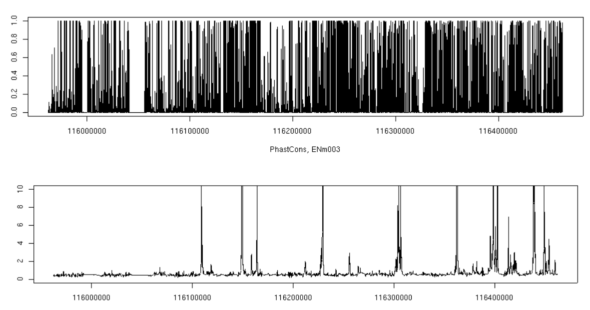
University of Washington

ABSTRACT

It has long been hypothesized that the human and other large genomes are organized into higher-order (i.e., greater than gene-sized) functional domains. Recent technological advances have enabled the rapid emergence of large-scale biological data sets comprising specific functional variables (e.g., transcription, histone modifications, etc.) sampled in a nearly continuous fashion across the genome. Here we develop novel approaches based on wavelet analysis for the discovery of "domain-level" behavior in fine scale functional genomic data, and for correlating apparently disparate functional data types collected at different resolutions and scales. We apply this approach to a variety of continuously sampled data types from the NHGRI ENCODE project. The results highlight an analytical framework which may be applied broadly to other complex genomes.

1 Introduction: continuous genomic data and *scale*

A wide variety of nearly continuous genomic data is now available in the approximately 1% (30Mb) of the human genome making up the ENCODE regions. Below, PhastCons conservation score (top), and DNaseI sensitivity (bottom, data from University of Washington) in the 500kb ENCODE region ENm003.

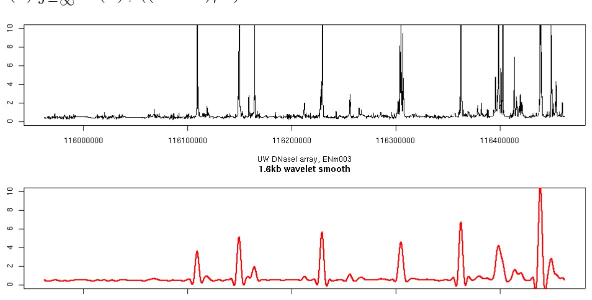


Questions we want to address focus on the issue of *scale*. We want to

1. compare and correlate datasets collected and presented on widely disparate scales; and
2. uncover trends in individual datasets at very large scales.

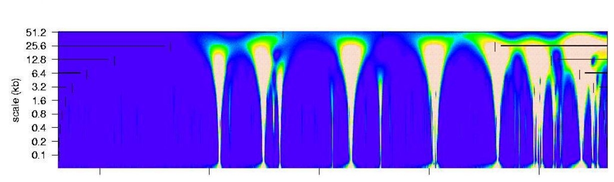
2 Methods: wavelets

Wavelets are a mathematical tool for capturing local behavior in data at any given scale. A wavelet is generally a function $\psi(t)$ such that $\int \psi = 0$, $\int |\psi|^2 = 1$. Given a function $X(t)$ to analyze, the wavelet coefficient $W(a, T)$ at scale a and time (genomic position) T is given by $W(a, T) = 1/\sqrt{(a)} \int_{-\infty}^{\infty} X(u)\psi((u - T)/a)\, du$.



Top: DNaseI sensitivity data. Bottom: the "wavelet smooth" approximation at the 1.6kb scale.

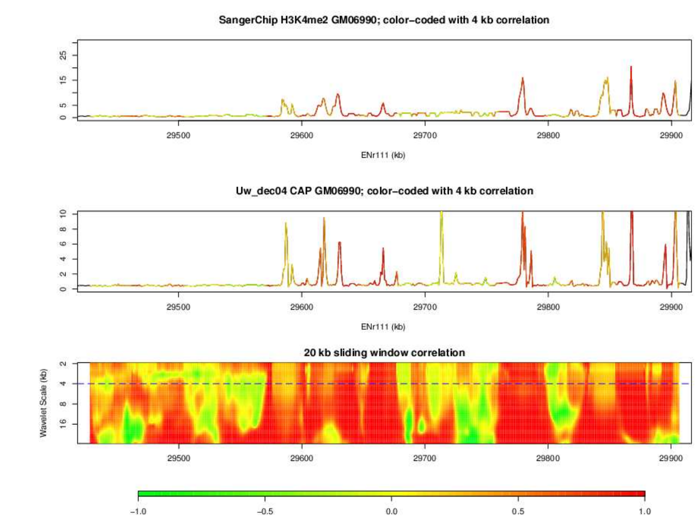
The wavelet *scalogram* is used to visualize trends in the data at a continuum of scales (y-axis).



3 Methods and results

Correlating disparate datasets using wavelets

We use wavelets to normalize disparate datasets to a common set of scales, then compute local correlations on a scale-by-scale basis.

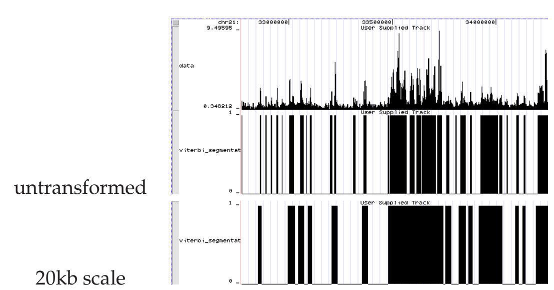


Each point in the correlation heatmap (bottom) represents the strength of the local correlation (red = positive, green = negative) at a particular scale. The raw data being correlated here is histone modification strength for H3K4me2 (data from Sanger Institute), which appears on top, and DNaseI sensitivity, bottom, in the 500kb ENCODE region ENr111. The raw data are color-encoded with the local correlations at the 4kb scale. The broad swaths of red indicate generally very high correlation between these two datasets.

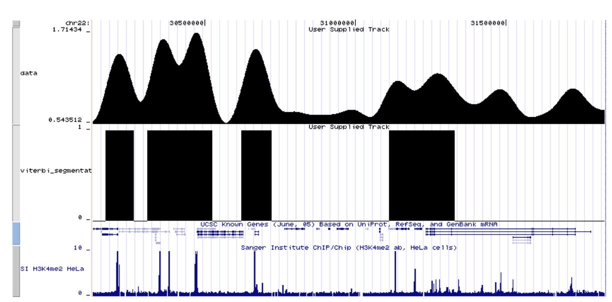
Generally, we observe a very high degree of correlation between DNaseI sensitivity and activating histone modifications H3K4me1, H3K4me2, H3K4me3, H3ac and H4ac (all from Sanger Institute). Against expectations, however, strongest correlations are with H3K4me2, H3K4me3, and weakest for H3ac, H4ac.

Segmentations of the ENCODE regions using HMMs and wavelets

We divide regions of the genome into "active" and "inactive" states using Hidden Markov Models (HMMs). By applying to wavelet-transformed data, we obtain scale-specific results, resulting in nested segmentations.

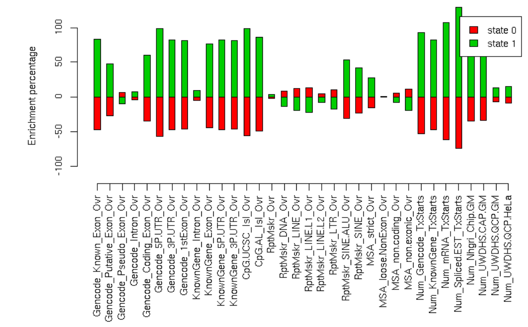


Above, segmentation based on Sanger H3K4me1 histone modification data. Below is a UCSC browser screen shot showing a 2-state segmentation (with the "active" state in the second row represented by black bars) of the 64kb wavelet smooth of H3K4me2, for the 1.7Mb ENCODE region ENm004. The smoothed data is plotted in the top track. The raw data appears in the final track. ENCODE-wide, the active state comprises 77 segments, with a median segment length of 109kb.



Results: Functional domain definition using single and multiple datatypes

Below is a summary plot of the enrichment or depletion of various genomic elements in each of the two states of the 64kb Sanger 2-state segmentation, measured ENCODE-wide against what would be expected at random.



We next computed a 2-state, simultaneous HMM segmentation of four datasets: TR50 replication timing, and loess-smoothed versions of RNA expression levels (data from Affymetrix), H3K4me2 and H3K27me3. The UCSC browser screen shot below, for ENm004, includes the smoothed datasets (top), followed by the segmentation, with the active state marked with black bars, and then raw data for the RNA, H3K27me3, and H3K4me2 datasets.

ENCODE-wide, the "active" state comprises 85 segments, with a median segment length of 94kb, covering approximately 41% of ENCODE (59% inactive).

At bottom is the enrichment plot as above, showing generally that using the Sanger data alone is sufficient to drive most of the stratification.