Higher-order functional domains in the human ENCODE regions

ROBERT THURMAN[a,b], NATHAN DAY[c], WILLIAM S. NOBLE[b], AND JOHN A. STAMATOYANNOPOULOS[b]

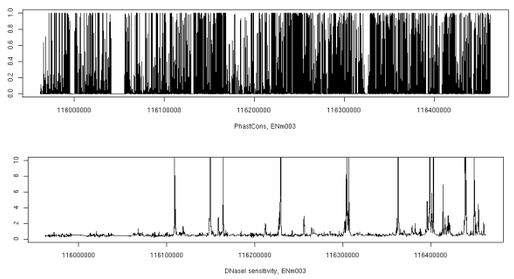
[a]Division of Medical Genetics, [b]Department of Genome Sciences, and [c] Department of Computer Science and Engineering, University of Washington, Seattle, WA, USA.

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. Under the ENCODE project, a number of functional experimental datatypes (e.g., histone modifications, transcription, etc.) have been measured in a nearly continuous fashion across the genome. We developed a novel approach combining wavelet analysis and Hidden Markov Models for unbiased discovery of "domain-level" behavior derived from these high-resolution functional genomic data. We find that higher order patterns in histone modifications, transcription, and DNA replication timing are generally concordant, and can be used to define discrete "active" and "inactive" functional domains, ranging from 15-500kb in size. Active and inactive domains differ markedly from one another with respect to annotated genomic features including gene content, CpG islands, and the spectrum of repetitive elements. One striking feature of this domain map is the degree to which different genomic territories highlighted by integrating multiple functional data types reflect an organization that cannot be readily predicted from the distribution of non-coding evolutionary conservation. The results collectively provide new insights into the functional landscape of the human genome.

1 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) in the 500kb ENCODE region ENm003.

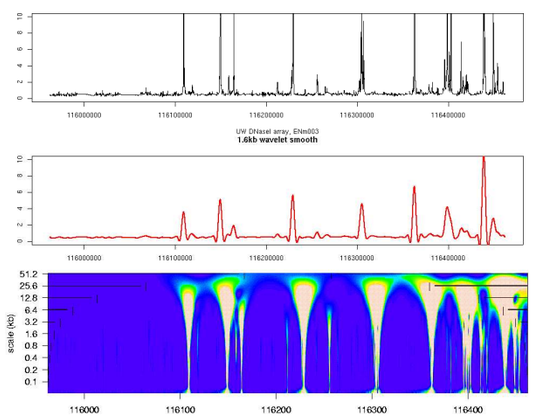


Questions we want to address focus on the issue of *scale*. We want to

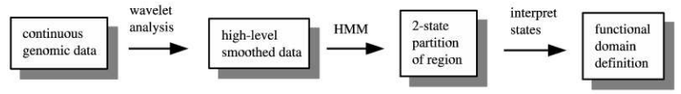
1. normalize datasets collected and presented on widely disparate scales; and
2. elucidate trends in individual and combined datasets at very large scales.

2 Wavelets

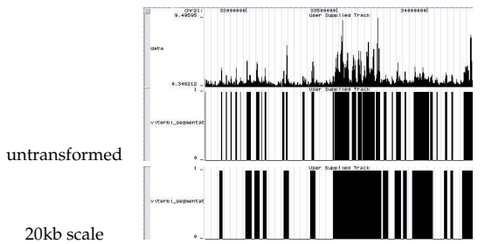
Wavelets provide a mathematical framework for capturing local behavior in data at any given scale. A wavelet is generally a function $\psi(t)$ such that $\int \psi = 0$, $\int |\psi|^2 = 1$. Given a function $X(t)$ to analyze, the wavelet coefficient $W(a,T)$ at scale a and time (genomic position) T is given by $W(a,T) = 1/\sqrt{(a)} \int_{-\infty}^{\infty} X(u)\psi((u-T)/a)\, du$.
At top, below, DNaseI sensitivity data. Beneath that is the "wavelet smooth" approximation at the 1.6kb scale. At bottom is the wavelet *scalogram*, used to visualize trends in the data at a continuum of scales (y-axis).



3 Segmenting ENCODE regions using HMMs and wavelets



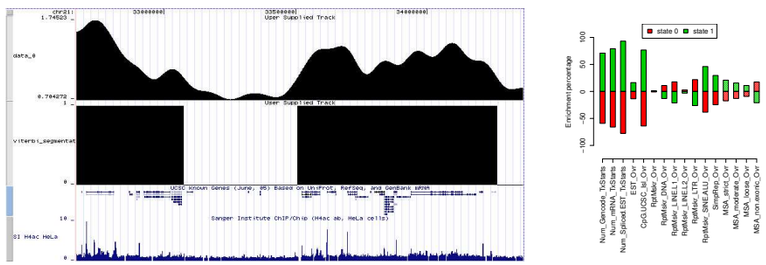
We divide regions of the genome into "active" and "inactive" states using Hidden Markov Models (HMMs). By applying HMMs to wavelet-transformed data we obtain scale-specific results, with larger scales producing coarser segmentations.



Segmentations based on Sanger H3K4me1 histone modification data.

Single-track segmentations

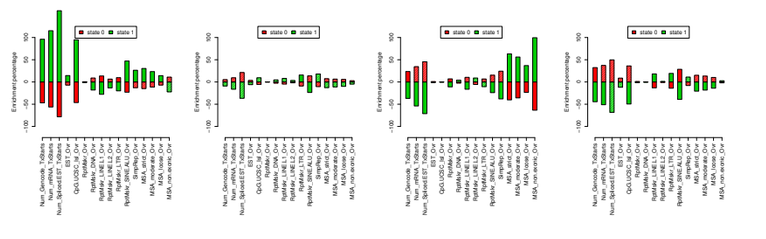
We computed 2-state segmentations based on 5 different ENCODE-wide continuous datasets: RNA transcription levels (Affymetrix), DNA replication timing (TR50, UVa), the activating histone mark H4ac (Sanger), the repressive mark H3K27me3 (UCSD), and conserved non-coding sequence density (ENCODE MSA group).



Above, left shows the segmentation (with the "active" state in the second row represented by black bars) of the 64kb wavelet smooth of H4ac for the 1.7Mb ENCODE region ENm005. The smoothed data is plotted in the top track. The raw data appears in the final track. ENCODE-wide, the active state comprises 55 segments, with a median segment length of 184kb. The figure on the right plots the enrichment of various genomic variables in the active (state 1, green) and inactive (state 0, red) states. Enrichment values significant at $p < 0.01$ are represented with solid bars as opposed to shaded. Below is table showing the concordance of each pair of the five segmentations, measured as the percentage of bases whose state assignments agree.

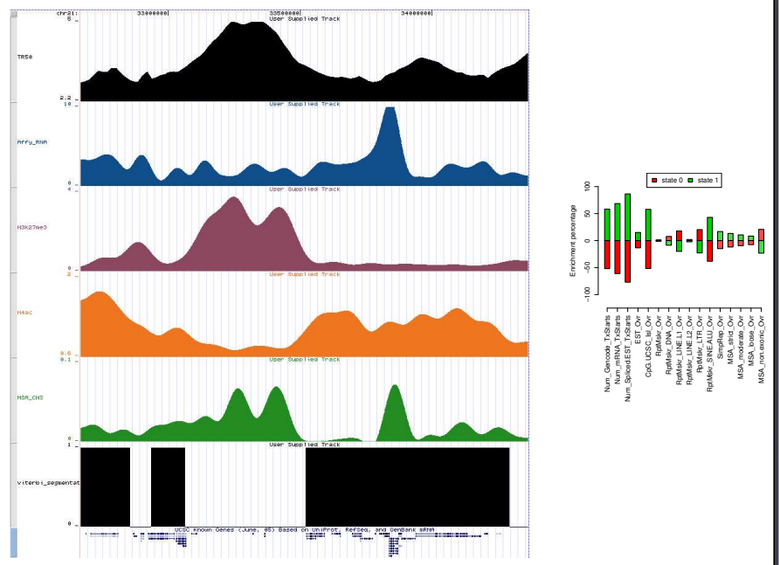
	H4ac	H3K27me3	Affy RNA	DNA repl (TR50)	CNS
H4ac	-	63%	74%	75%	61%
H3K27me3	63%	-	49%	70%	54%
Affy RNA	74%	49%	-	61%	57%
DNA repl (TR50)	75%	70%	61%	-	57%
CNS	61%	54%	57%	57%	-

Below are segmentation enrichment plots for (left to right) RNA, H3K27me3, CNS, and DNA replication.



Functional domain definition using multiple datatypes simultaneously

We next computed a 2-state segmentation using the five datasets simultaneously as input to the HMM. The UCSC browser screen shot below, for ENm005, includes the smoothed datasets (top), followed by the segmentation.
ENCODE-wide, the "active" state for this segmentation comprises 53 segments, with a median segment length of 180kb, covering approximately 13.8Mb, or 46% of ENCODE (54% inactive).
At right is the enrichment plot as before, showing that the multiple track segmentation reproduces some of the strong patterns of the more highly enriched single-track segmentations.



Multiple-track segmentation provides robust domain definition

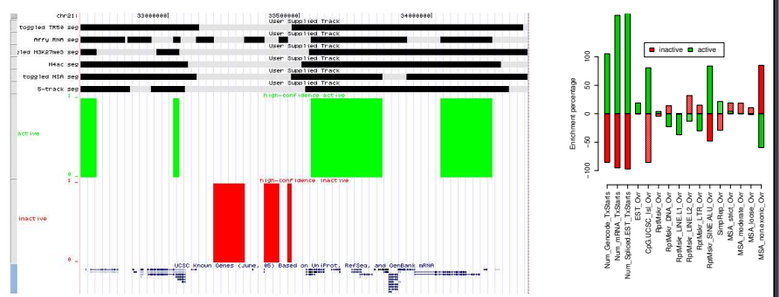
Below is the concordance of the five-track segmentation with the single-track segmentations of each of the five constituent data-types.

	H4ac	H3K27me3	Affy RNA	DNA repl (TR50)	CNS
Five-track	89%	72%	70%	82%	61%

The five-track segmentation shows an overall higher degree of concordance with each of the constituent single-track segmentations than exists between the single-track segmentations themselves. Given the enrichment profile above, the five track segmentation thus provides a robust summary of each of the individual data types.

High-concordance functional and non-functional regions

We finally define regions of agreement between all six segmentations (five-track, plus each of the five single-track segmentations) as high-concordance active and inactive regions. These high-concordance regions are comprised of 48 active segments (green bars, left), covering 5.2Mb, and 25 inactive segments (red), covering 2.2Mb, for a total coverage of approximately 25% of ENCODE. The enrichment profile for these segments, right, is significantly enhanced versus the individual segmentation results.



High-concordance segments in ENm005. At top, left are active (black) and inactive (gray) segments as defined by the single-track and five-track segmentations. Green and red bars below that represent the intersection of the active and inactive segments, respectively. At right, the enrichment profile for the high-concordance regions.