

# Thoughts on ENCODE Annotations

Mark Gerstein

# Genomic annotations

## Introduction

The ENCODE Project provides a set of candidate genomic regions that can serve as predictions for further investigation. This page provides links to download a set of candidate genomic regions as well as a list of publications that contain additional data.

## Candidate genomic regions

- Gene expression matrix over ENCODE2 cell lines (~60 cell lines in total) in GENCODE 19 [Download data | Download methods]
- GENCODE v19 TSS list stratified by Fantom5 CAGE data [View README]
  - Strict CAGE clusters [Download]
  - Robust CAGE clusters [Download]
  - Permissive CAGE clusters [Download]
- Candidate enhancers based on DNase hypersensitivity and H3K27ac and annotated with TF-ChIP peaks as well as candidate promoters annotated with TF-ChIP peaks. [Visualize data | Download methods]
  - Distal DNase peaks [Download]
  - Proximal DNase peaks [Download]
  - H3K27ac annotations [Download]
  - Distal TF binding sites [Download]
  - Proximal TF binding sites [Download]

## Additional annotations

Papers previously published by the ENCODE Consortium contain data files that include additional genomic annotations. Search for all publications with ENCODE element data

### Peaks

Peaks are enriched regions of the genome corresponding to either sites of transcription factor binding or DNase hypersensitivity identified during various functional genomic assays. In this section, we provide a list of peaks in various cell lines using both DNase-Seq and ChIP-Seq assays. [View publications.](#)

### RNAs

RNA represents the direct readout of the genetic information encoded by genomes and a significant proportion of a cell's regulatory capabilities are focused on its synthesis, processing, transport, modification and translation. A catalogue of the RNA species made inside the cell and the amount of RNA from each of these loci across various cell lines is provided in this section. [View publications.](#)

### Promoters

The promoter is the region proximal to the transcription start site of a gene that regulates its transcription using transcription factor binding sites. These transcription factors recruit RNA polymerase after binding to the promoter and initiate transcription of the gene. [View publications.](#)

**Simplified**

**Comprehensive**  
(published annotation, mostly  
in '12 & '14 rollouts)

# “Comprehensive” Functional Genomics Annotations

## A) PEAKS

1. DNase peaks at the UCSC genome browser {on many cell lines}
2. The regulation track at the UCSC genome browser, with compilation of TF ChIP-seq peaks from uniform processing (individual peaks are annotated with TF and cell line)
3. Blacklist Regions

## B) PROMOTERS

Annotated GENCODE TSSes (also, TSSes with FANTOM CAGE support)

## C) ENHANCERS (Supervised)

## D) UNSUPERVISED SEGMENTATIONS, INCLUDING ENHANCERS

ChromHMM, SegWay, HiHMM....

## E) HOT/LOT REGIONS

## F) CONNECTIVITY

1. Enhancer-target gene connection
2. TF-target network connectivity
3. TADs: Topologically Associated Domain

## G) MOTIFS

for TF binding

## H) RNA

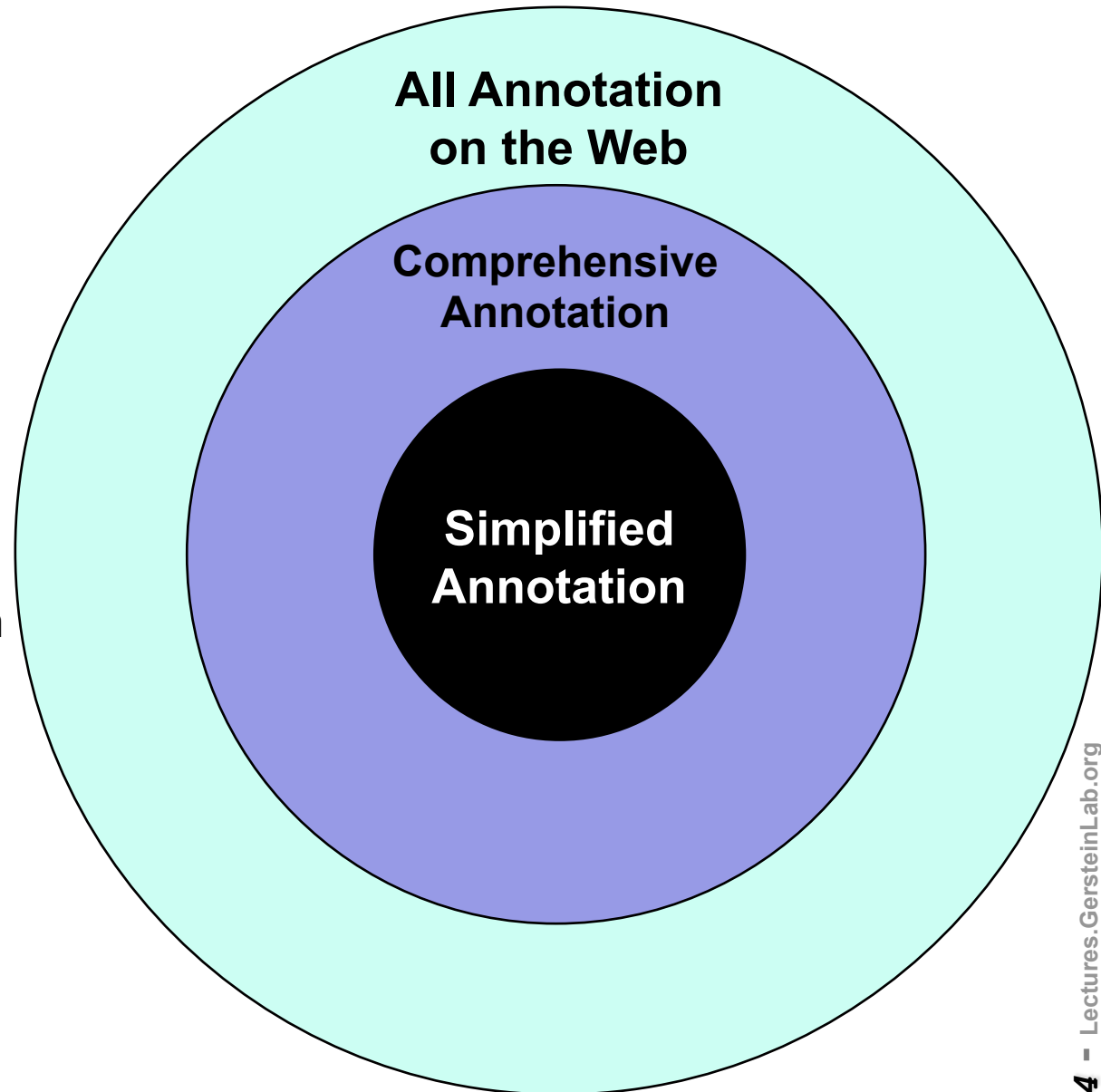
1. A matrix of expression data of known genes (or exons) for protein-coding genes & known ncRNAs {on many cell lines}
2. Novel RNA contigs track, i.e., possible novel transcripts (ie Transcriptionally Active Regions or TARs)
3. Novel junctions

## I) OTHER

1. List of Allelic SNPs & Regions
2. Models

## “App Store” for Annotation v Defaults: Encourage innovation & remain organized

- Simplified, "monolithic" annotation **deliberately does not include** everything
- Want to **encourage innovation** by many both inside & outside consortium -- yet still have organization. Want to feature not squelch work of others.
- **Metaphor:**  
**App Store** v Default Apps v everything on web
- **Need to develop a process** for getting into the ENCODE "app store" – eg specs, being published, give a presentation on the AWG, &c
- How does ENCODE interact w/ individuals creating annotation from ENCODE data, other annotation resources (eg ENSEMBL), other consortia developing annotation (eg GTEx)



# Additional Issues to Consider Beyond What's in Simplified Annotations

- How to represent uncertainty in the boundaries
  - Overlapping & conditional annotations (e.g. this UTR goes with this promotor & enhancer)
- How to represent connectivity & large-scale structures ?
- How to represent on personal genomes (including cancer genomes)
  - Personal annotation on NA12878
  - Mapping to the personal genome
- Relationship to eQTLs & allelic SNPs

Sources of Annotation: Comparative & Functional

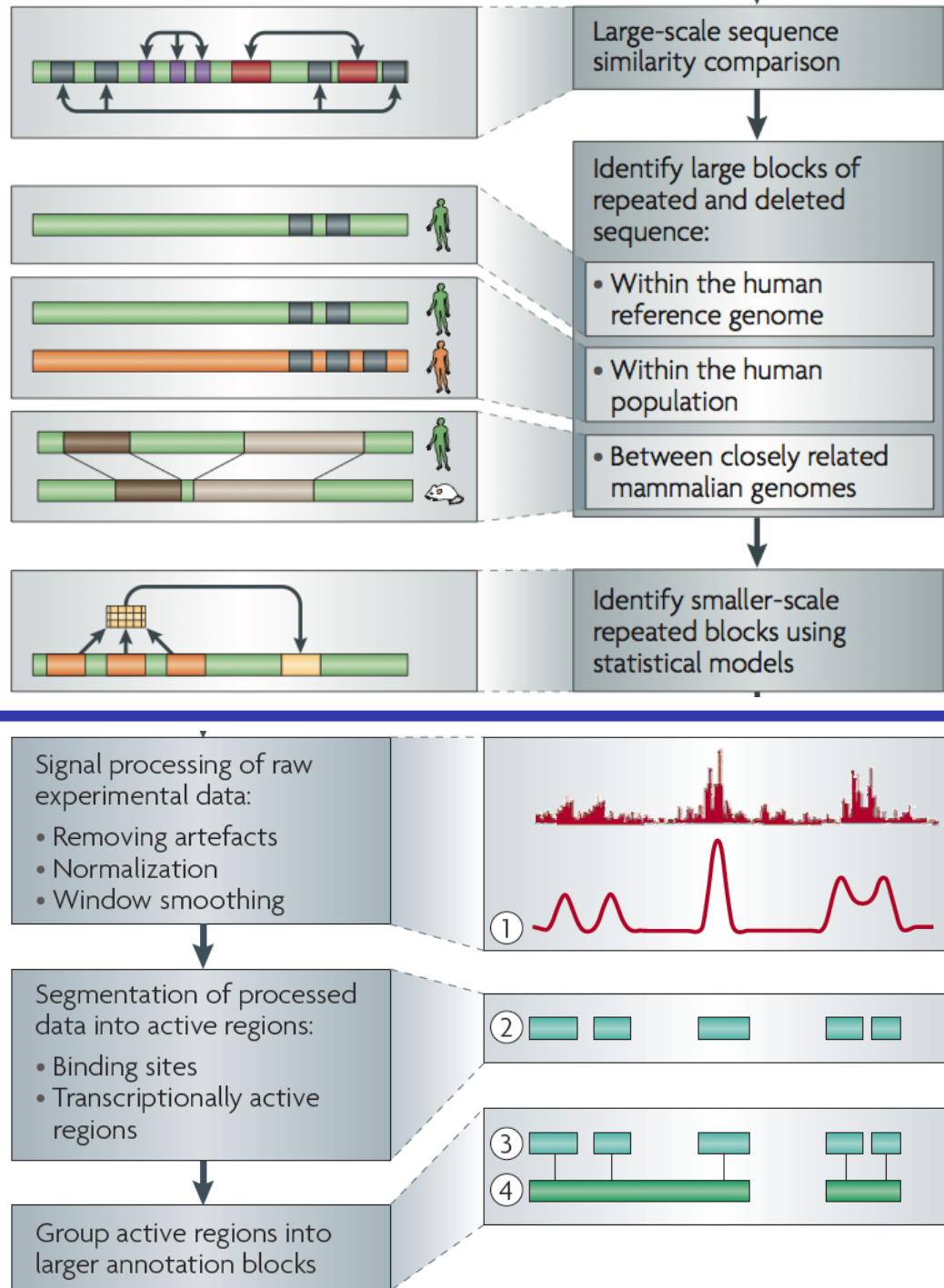
# Integrating Conservation

Across organisms & just in humans

Not trivial to represent (ie related to short size of elements)

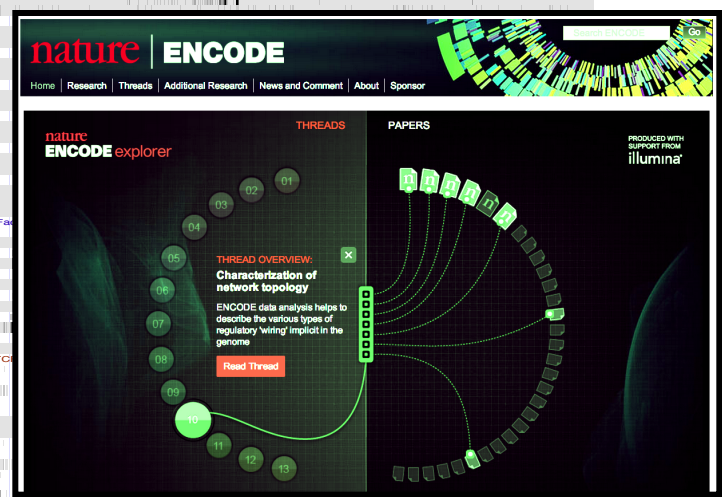
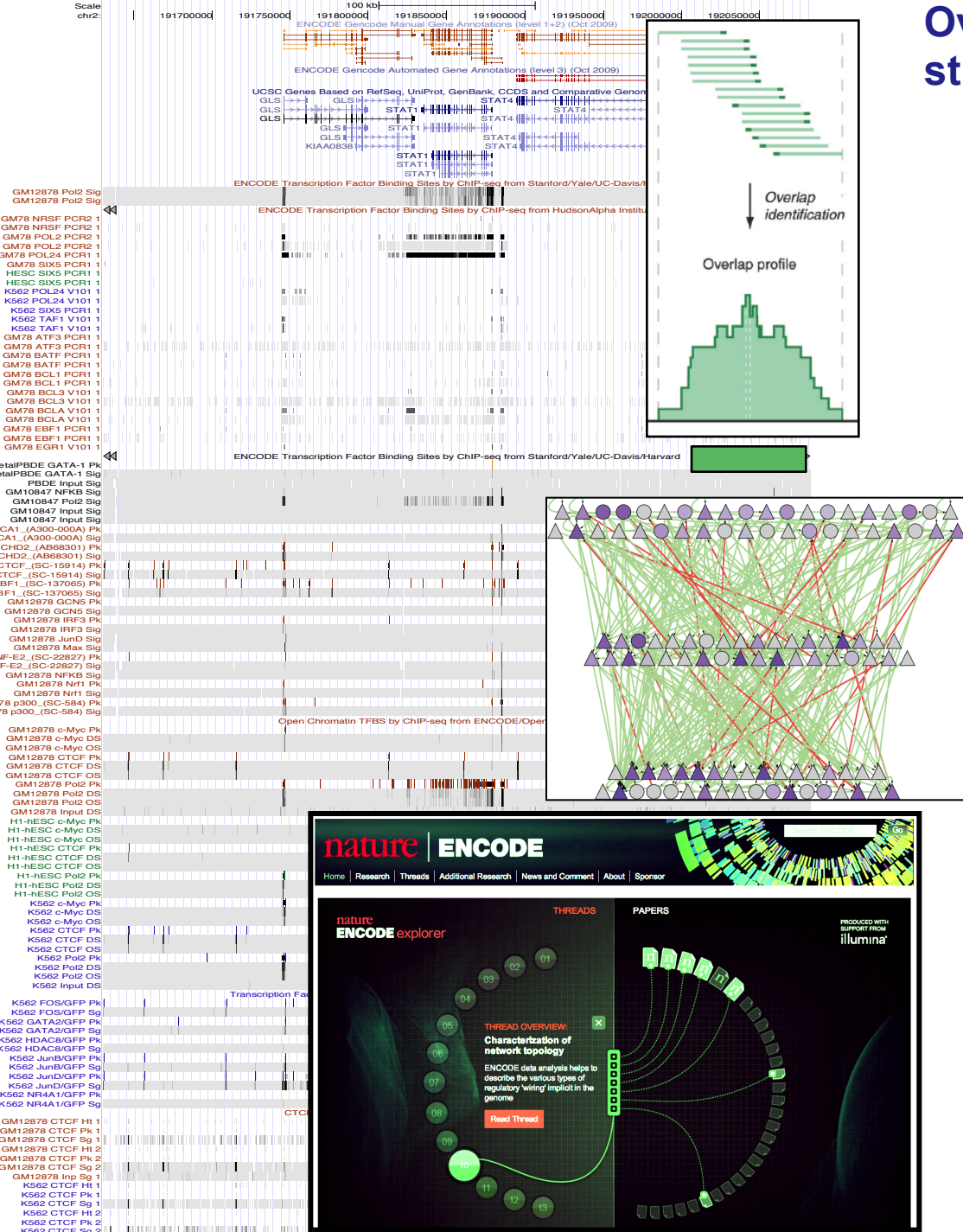
Many representations: Motifs, conserved "peaks", CpGs

What "critics" often bring up



# Overall ENCODE presentation as a structured hierarchy

- Raw data (reads) at the bottom
- Progressive Processed Summaries
  - Signals (e.g. representing the degree to which DNA is bound by TFs)
  - (Simplified Annotation?)
  - Site locations
  - Reg. networks, chromatin states & stat. models
- Code & VMs
- At top are linked publications documenting everything, forming metadata
  - Preferable to separate online docs.



[PLOS CB 4:e1000158;  
Nature.com/encode + Nature 489: 208]