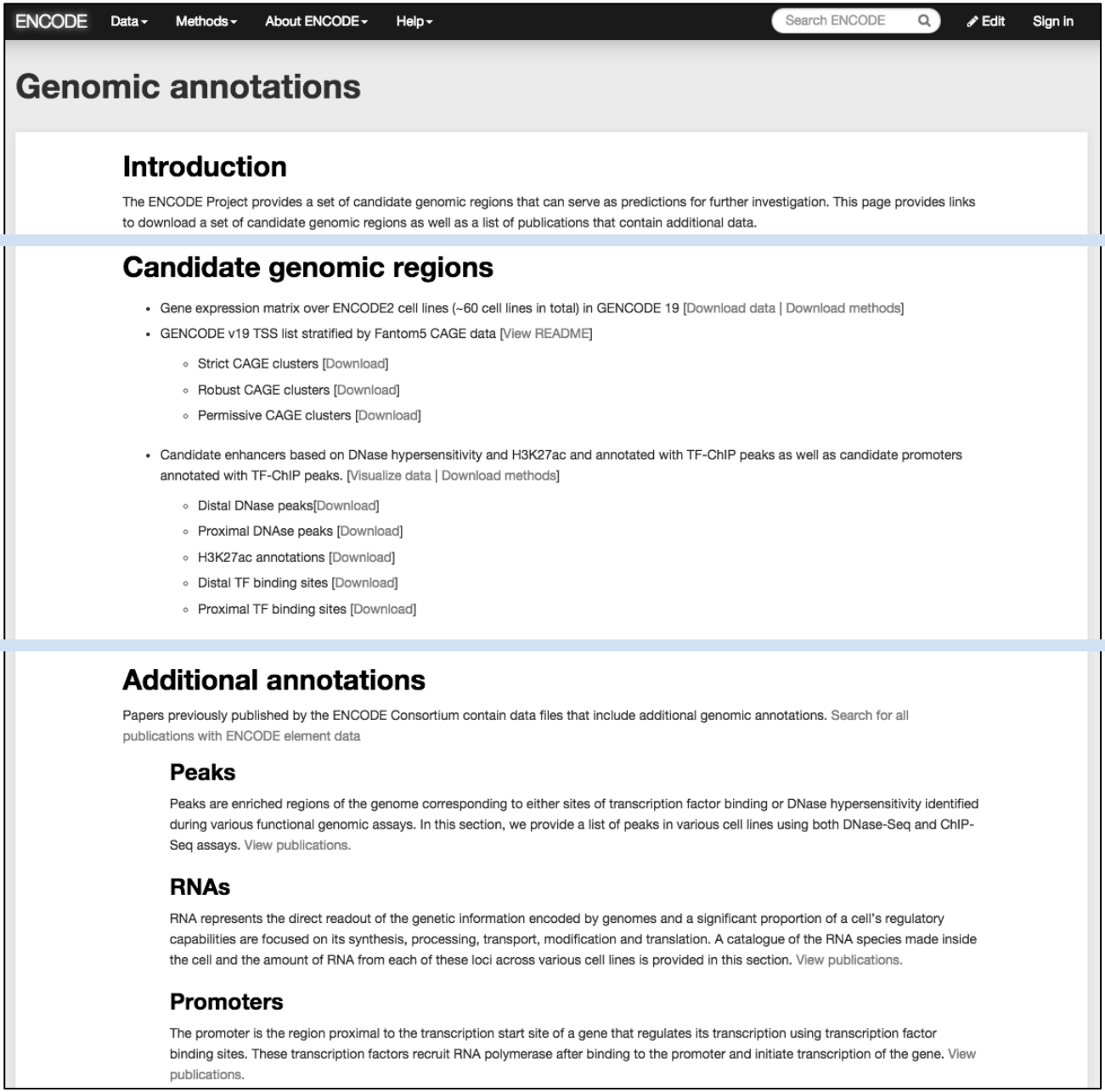
Thoughts on ENCODE Annotations

Mark Gerstein



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)

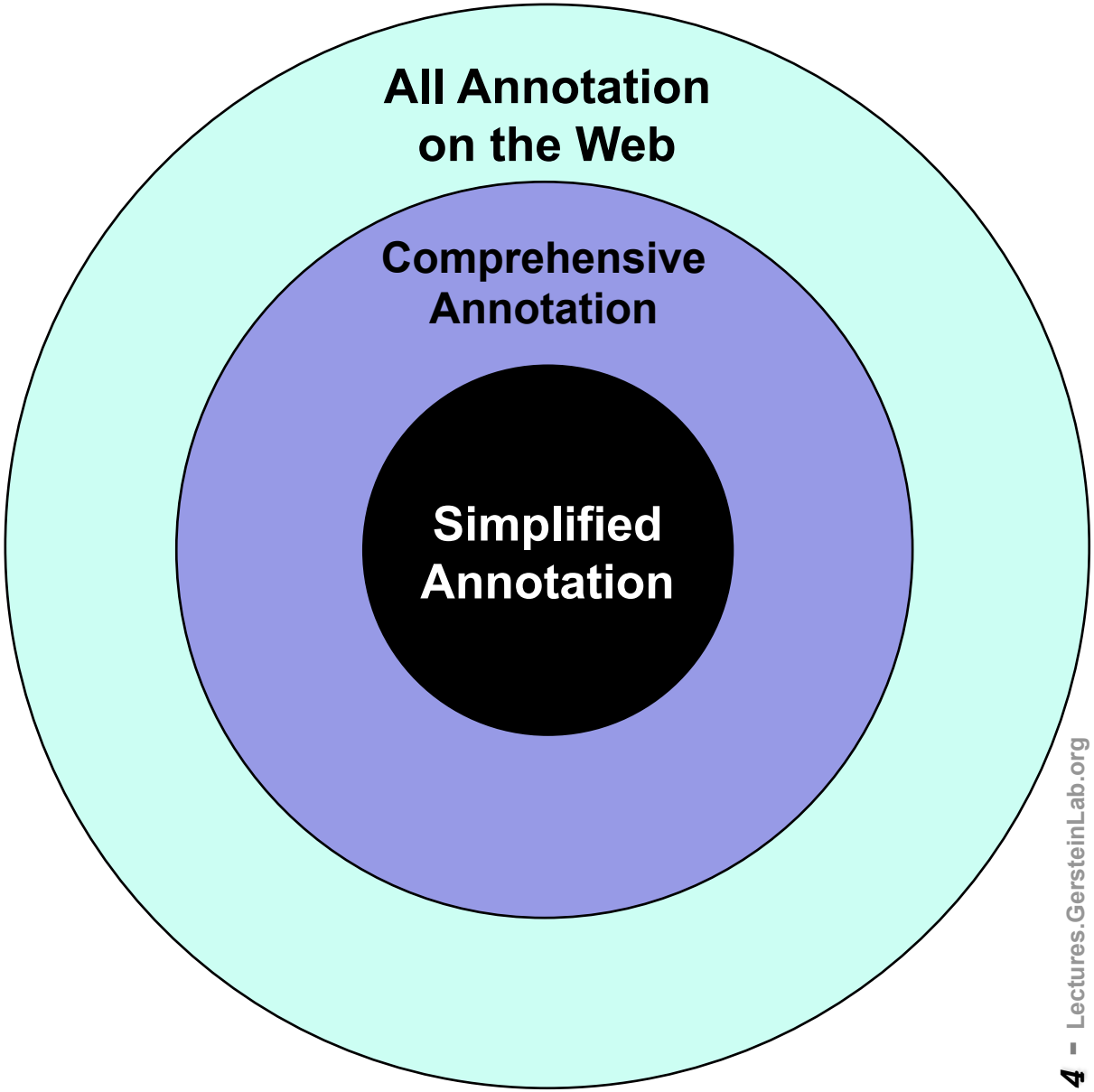
3. Novel junctions

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

[Nat. Rev. Genet. (2010) 11: 559]

Sources of Annotation: Comparative & Functional

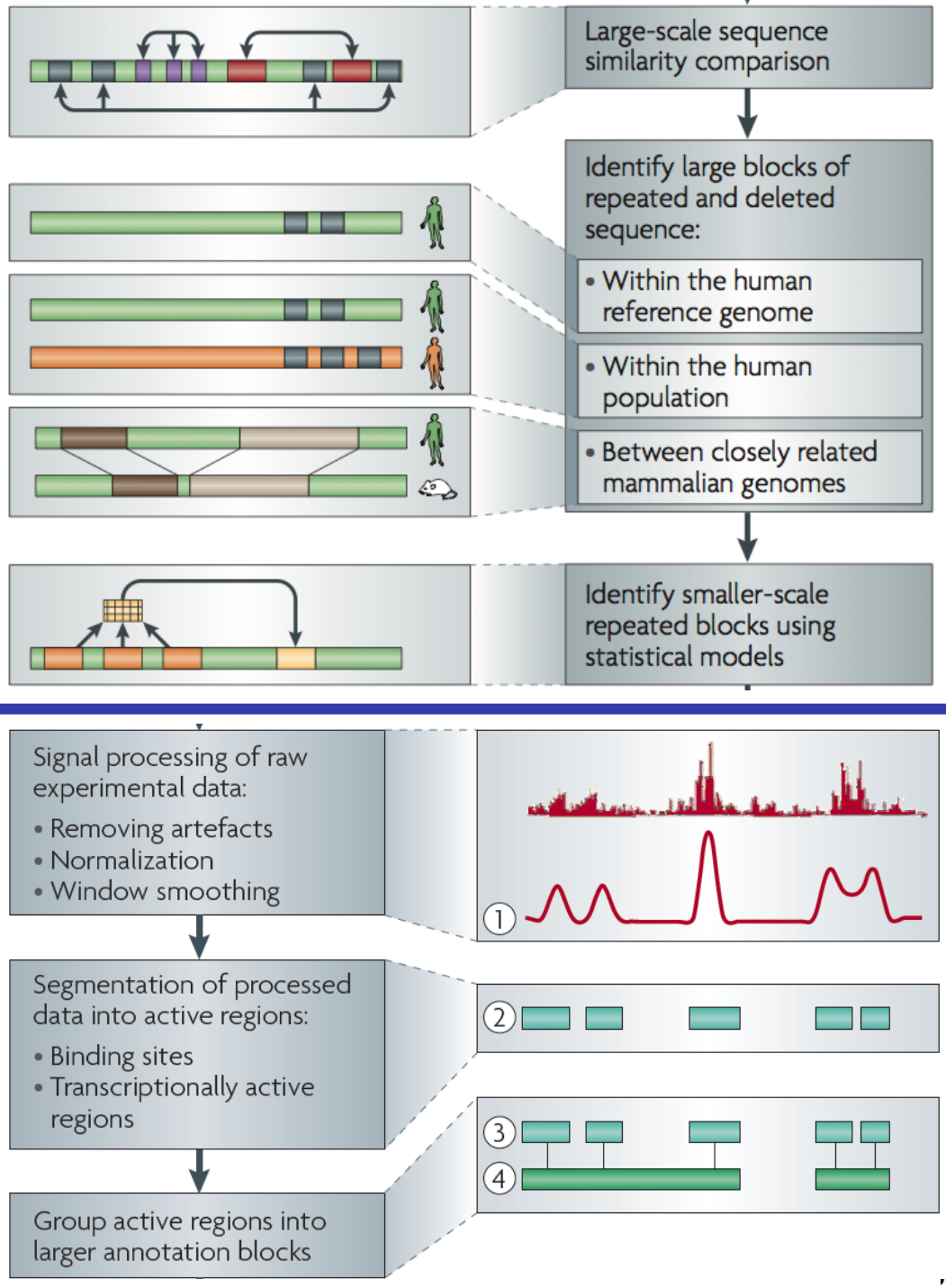
Integrating Conservation

Across organisms & just in humans

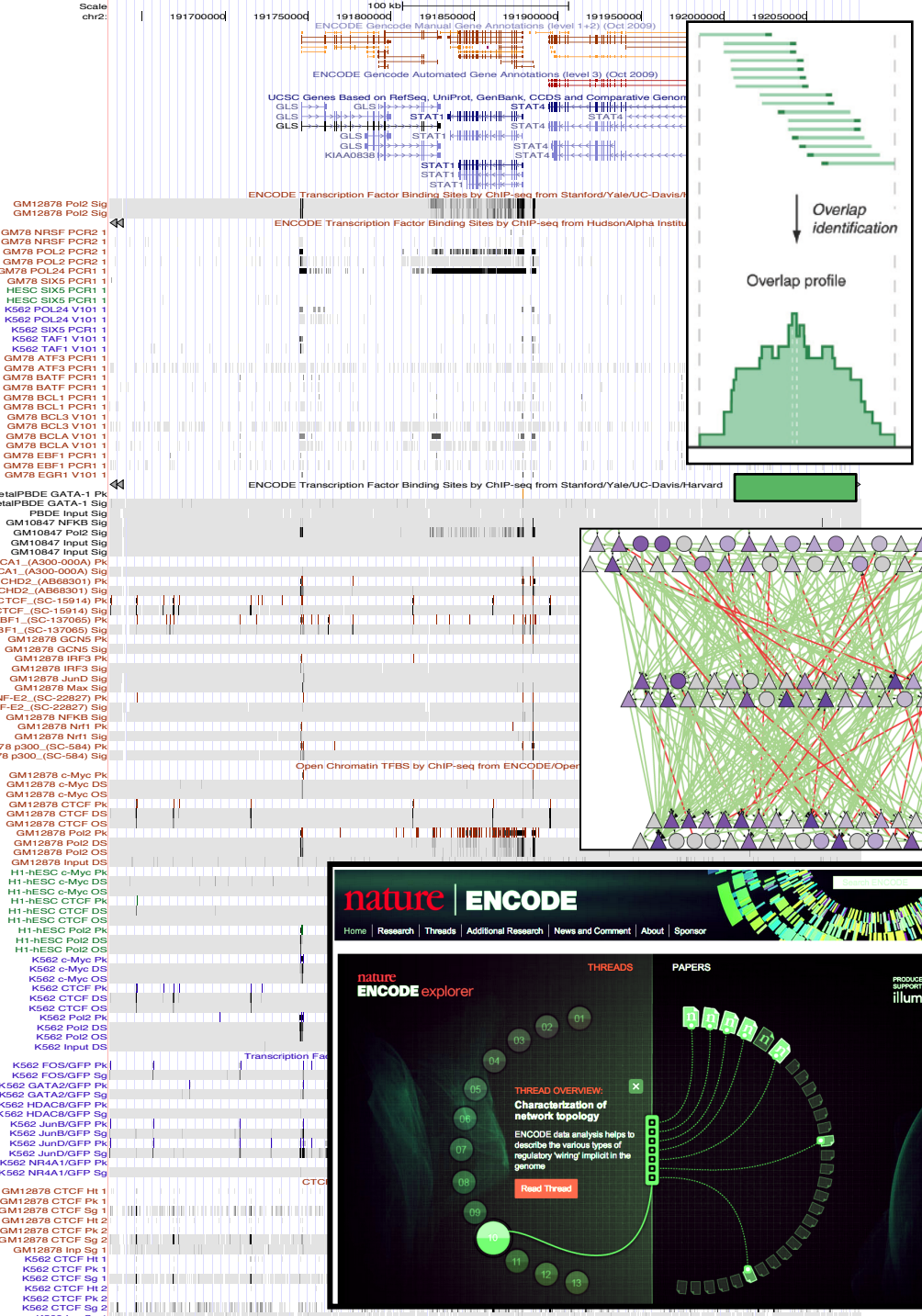
Not trivial to represent (ie related to short size of elements)

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]