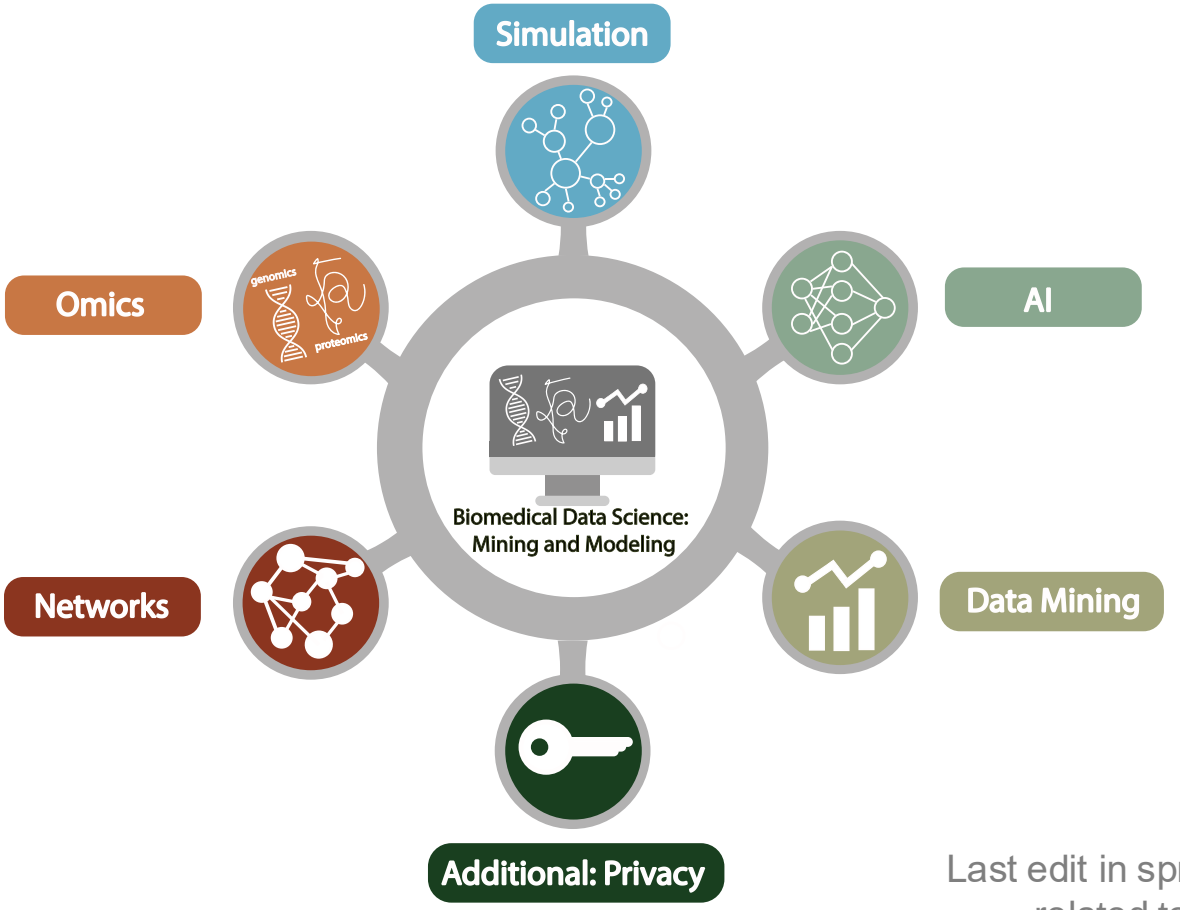
Biomedical Data Science (GersteinLab.org/courses/452)

Genome Annotation (Multi-omic Analyses) (25m7-part1)



Mark Gerstein
Yale U.

Last edit in spring '25. Just first half related to genome annotation. Reduced integration section. Now loosely related to 1st half of 2021's M7 [which has a video].

What is Annotation? (For Written Texts?)

No. 4356 April 25, 1953 NATURE

MOLECULAR STRUCTURE OF NUCLEIC ACIDS

A Structure for Deoxyribose Nucleic Acid

WE wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest.

A structure for nucleic acid has already been proposed by Pauling and Corey[1]. They kindly made their manuscript available to us in advance of publication. Their model consists of three intertwined chains, with the phosphates near the fibre axis, and the bases on the outside. In our opinion, this structure is unsatisfactory for two reasons: (1) We believe that the material which gives the X-ray diagrams is the salt, not the free acid. Without the acidic hydrogen atoms it is not clear what forces would hold the structure together, especially as the negatively charged phosphates near the axis will repel each other. (2) Some of the van der Waals distances appear to be too small.

Another three-chain structure has also been suggested by Fraser (in the press). In his model the phosphates are on the outside and the bases on the inside, linked together by hydrogen bonds. This structure as described is rather ill-defined, and for this reason we shall not comment on it.

We wish to put forward a radically different structure for the salt of deoxyribose nucleic acid. This structure has two helical chains each coiled round the same axis (see diagram). We have made the usual chemical assumptions, namely, that each chain consists of phosphate diester groups joining β-D-deoxyribofuranose residues with 3′,5′ linkages. The two chains (but not their bases) are related by a dyad perpendicular to the fibre

NATURE | VOL 409 | 15 FEBRUARY 2001 |

Initial sequencing and analysis of the human genome

International Human Genome Sequencing Consortium*

* *A partial list of authors appears on the opposite page. Affiliations are listed at the end of the paper.*

The human genome holds an extraordinary trove of information about human development, physiology, medicine and evolution. Here we report the results of an international collaboration to produce and make freely available a draft sequence of the human genome. We also present an initial analysis of the data, describing some of the insights that can be gleaned from the sequence.

The rediscovery of Mendel's laws of heredity in the opening weeks of the 20th century[1–3] sparked a scientific quest to understand the nature and content of genetic information that has propelled biology for the last hundred years. The scientific progress made falls naturally into four main phases, corresponding roughly to the four quarters of the century. The first established the cellular basis of heredity: the chromosomes. The second defined the molecular basis of heredity: the DNA double helix. The third unlocked the informational basis of heredity, with the discovery of the biological mechanism by which cells read the information contained in genes and with the invention of the recombinant DNA technologies of cloning and sequencing by which scientists can do the same.

The last quarter of a century has been marked by a relentless drive to decipher first genes and then entire genomes, spawning the field of genomics. The fruits of this work already include the genome sequences of 599 viruses and viroids, 205 naturally occurring plasmids, 185 organelles, 31 eubacteria, seven archaea, one fungus, two animals and one plant.

Here we report the results of a collaboration involving 20 groups from the United States, the United Kingdom, Japan, France, Germany and China to produce a draft sequence of the human genome. The draft genome sequence was generated from a physical map covering more than 96% of the euchromatic part of the human genome and, together with additional sequence in public databases, it covers about 94% of the human genome. The sequence was produced over a relatively short period, with coverage rising from about 10% to more than 90% over roughly fifteen months. The sequence data have been made available without restriction and updated daily throughout the project. The task ahead is to produce a finished sequence, by closing all gaps and resolving all ambiguities. Already about one billion bases are in final form and the task of bringing the vast majority of the sequence to this standard is now straightforward and should proceed rapidly.

coordinate regulation of the genes in the clusters.

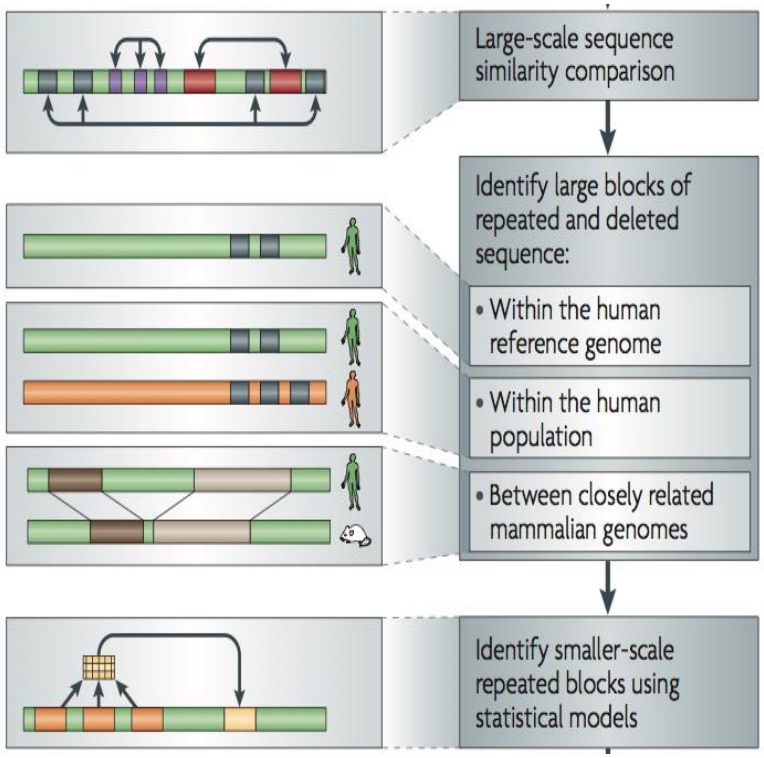
- There appear to be about 30,000–40,000 protein-coding genes in the human genome—only about twice as many as in worm or fly. However, the genes are more complex, with more alternative splicing generating a larger number of protein products.
- The full set of proteins (the 'proteome') encoded by the human genome is more complex than those of invertebrates. This is due in part to the presence of vertebrate-specific protein domains and motifs (an estimated 7% of the total), but more to the fact that vertebrates appear to have arranged pre-existing components into a richer collection of domain architectures.
- Hundreds of human genes appear likely to have resulted from horizontal transfer from bacteria at some point in the vertebrate lineage. Dozens of genes appear to have been derived from transposable elements.
- Although about half of the human genome derives from transposable elements, there has been a marked decline in the overall activity of such elements in the hominid lineage. DNA transposons appear to have become completely inactive and long-terminal repeat (LTR) retroposons may also have done so.
- The pericentromeric and subtelomeric regions of chromosomes are filled with large recent segmental duplications of sequence from elsewhere in the genome. Segmental duplication is much more frequent in humans than in yeast, fly or worm.
- Analysis of the organization of Alu elements explains the long-standing mystery of their surprising genomic distribution, and suggests that there may be strong selection in favour of preferential retention of Alu elements in GC-rich regions and that these 'selfish' elements may benefit their human hosts.
- The mutation rate is about twice as high in male as in female meiosis, showing that most mutation occurs in males.
- Cytogenetic analysis of the sequenced clones confirms suggestions that large GC-poor regions are strongly correlated with 'dark

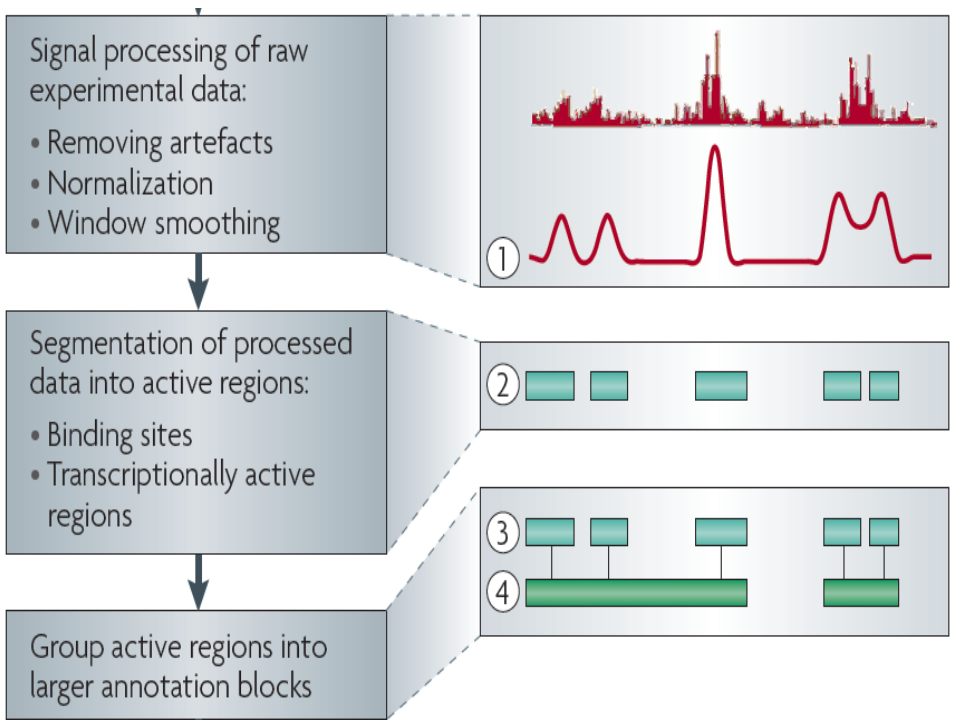
Non-coding Annotations: Overview

Features are often present on multiple ”scale” (eg elements and connected networks)

Sequence features, incl. **Conservation**

Functional Genomics
Chip-seq (Epigenome & seq. specific TF) and ncRNA & un-annotated transcription



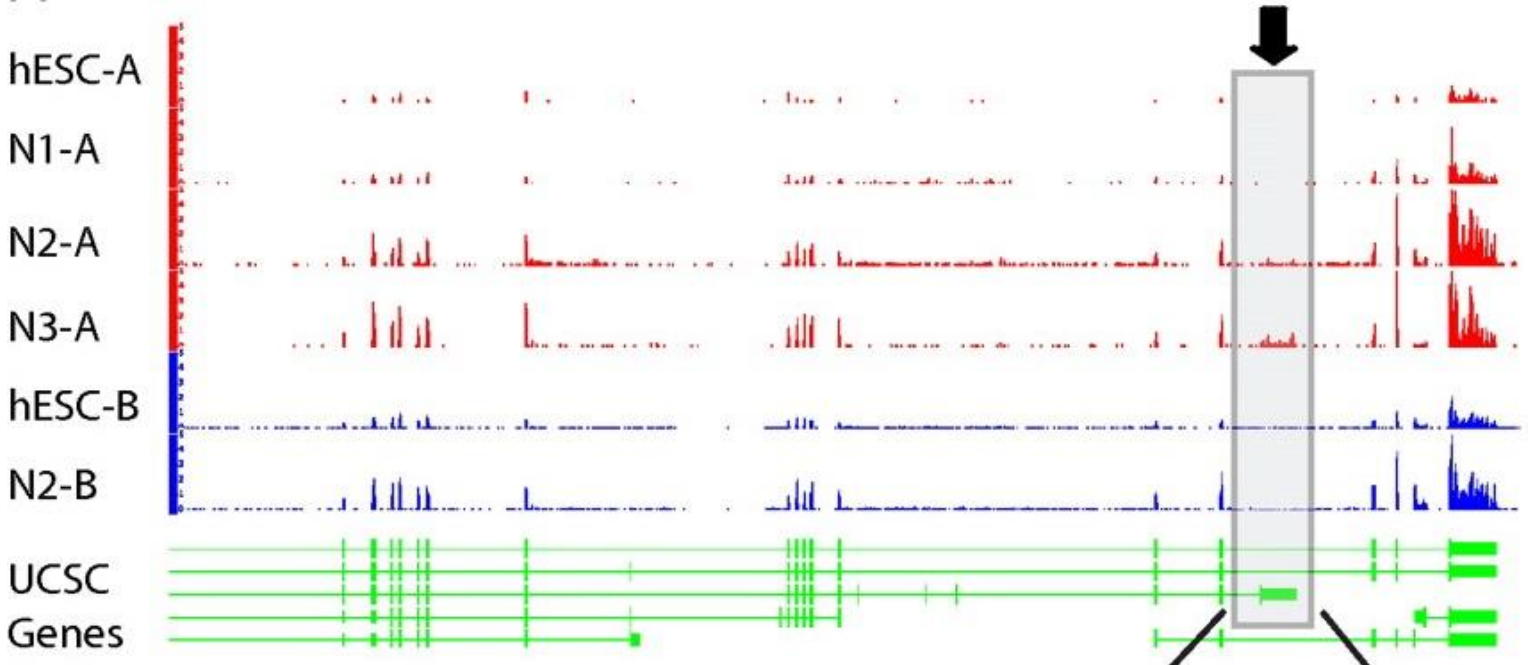


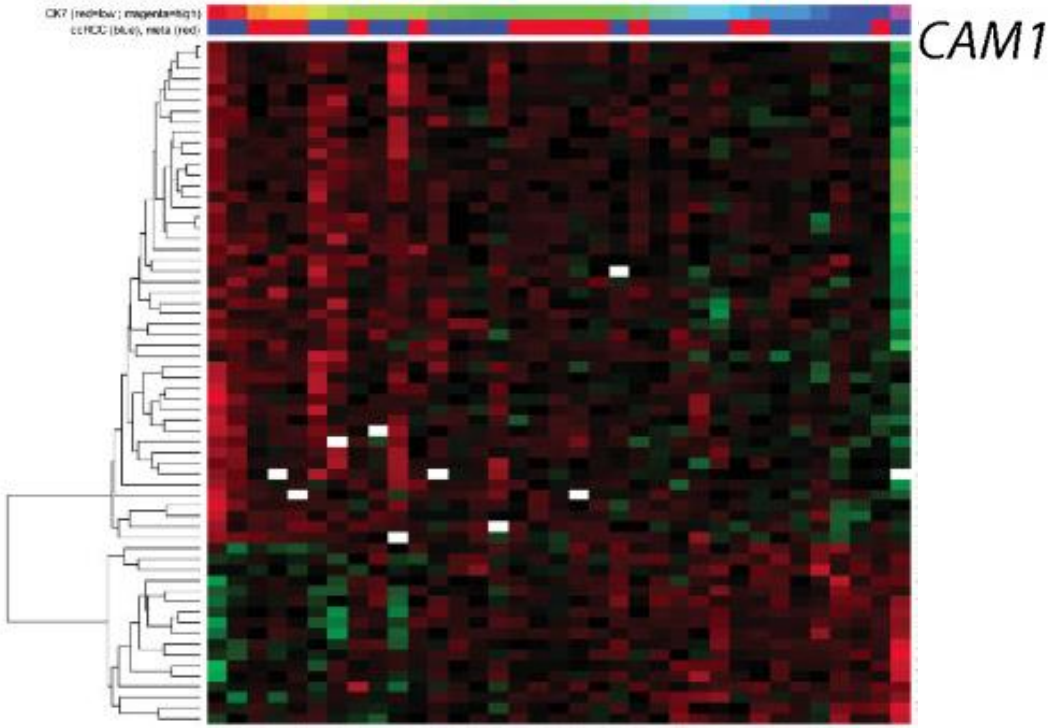
[*Nat. Rev. Genet.* (2010) 11: 559]

Outline

- **Part 1 : Generic Annotation (not related to an individual's variants)**
  - RNA-seq, Chip-seq
  - Integration
  - , Hi-C
- Part 2 : Annotation related to an individual's variants
  - ASE/ASB
  - GWAS & eQTL

RNA-seq & Chip-seq

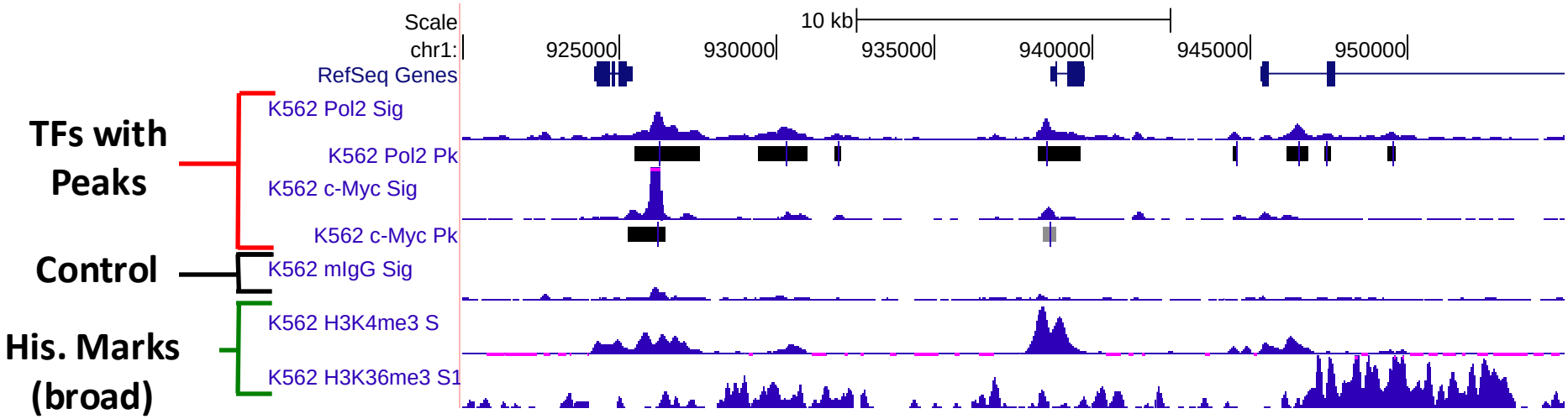




Information from RNA-seq: Avg. signal at exons & TARs (RPKMs)

[*PNAS* 4:107: 5254 ; *IJC* 123:569]

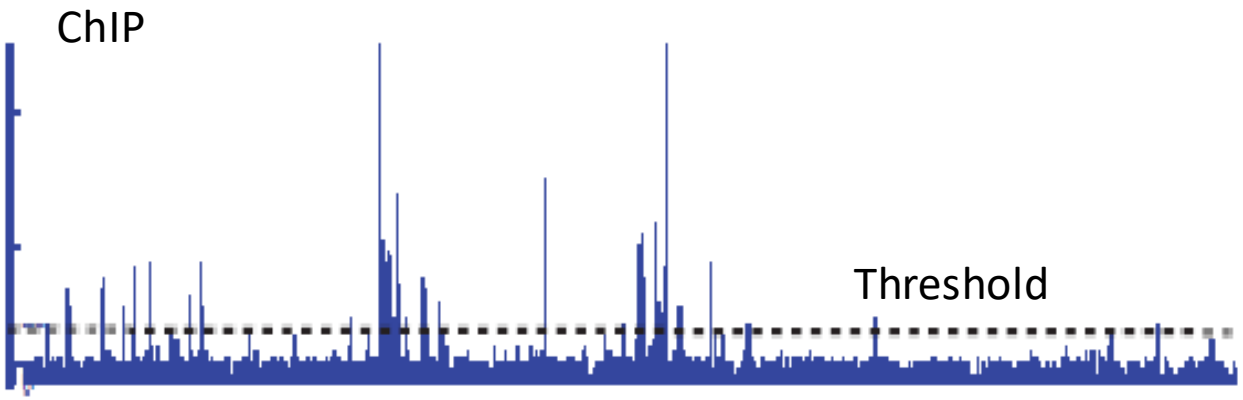
Information from Chip-seq



[*Science* 330: 1775
+ ENCODE Data Sources
TFs & Control: Yale
HMs: UW & Broad]

Summarizing the Signal: "Traditional" ChipSeq Peak Calling

- Generate & threshold the signal profile to identify candidate target regions
  - Simulation (PeakSeq),
  - Local window-based Poisson (MACS),
  - Fold change statistics (SPP)



Potential Targets →

- Score against the control

Significantly Enriched targets →

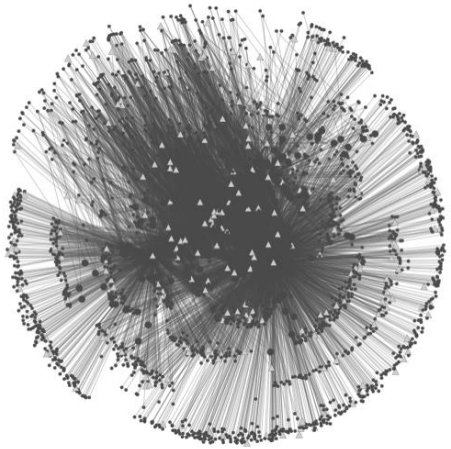
[Rozowsky et al. ('09) *Nat Biotech*]

Data Flow: peaks to proximal & distal networks

Peak Calling

Assigning TF binding sites to targets

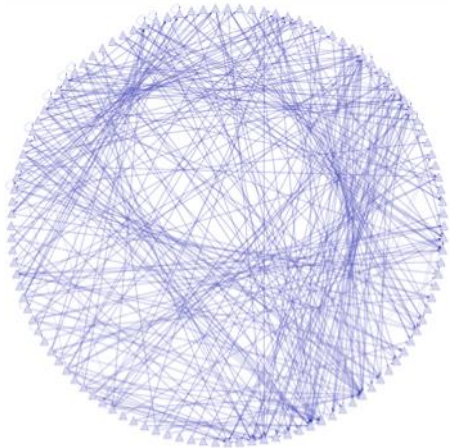
~500K Edges



Filtering high confidence edges & distal regulation

Based on stat. model combining
signal strength & location relative to typical binding

~26K Edges



TF

Potential Distal Edge

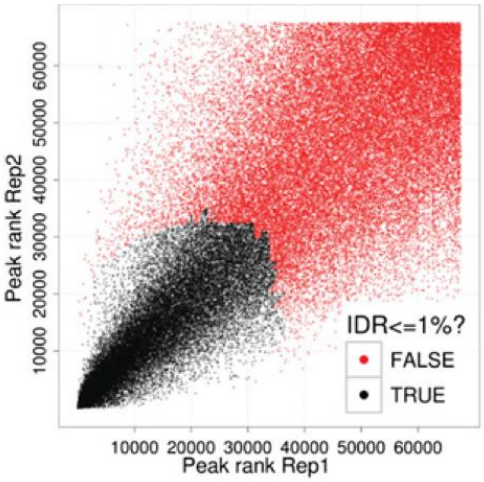
TF

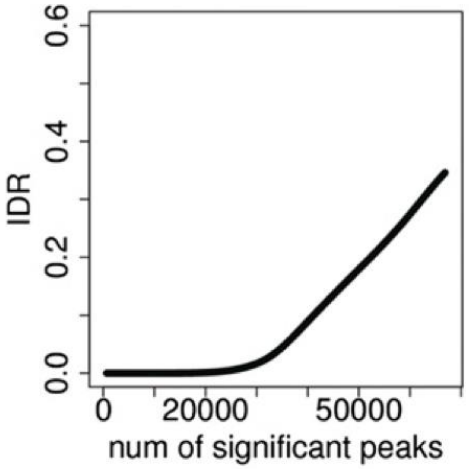
Strong Proximal Edge

[Cheng et al., *Bioinfo.* ('11);
Gerstein et al. *Nature* (in press, '12) ;
Yip et al., *GenomeBiology* (in press, '12)]

The irreproducible discovery rate (IDR)

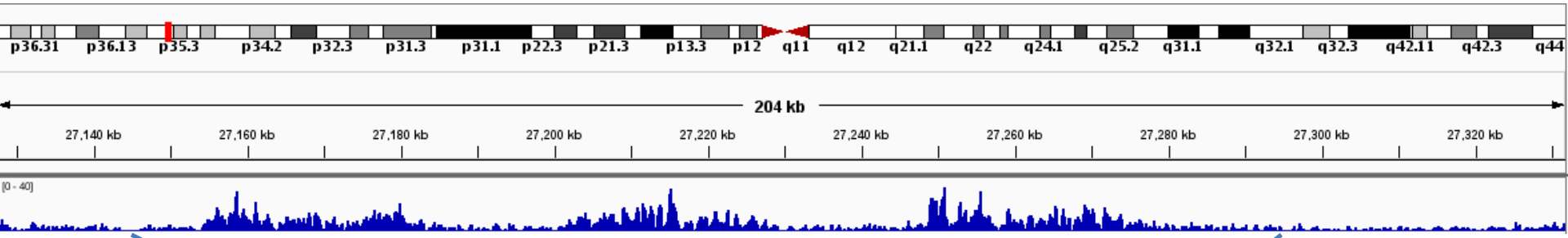
- Unified approach to measure the reproducibility of findings identified from replicate high-throughput experiments.
- Idea : call peaks with low cutoff and classify peaks as reproducible or not (bivariate rank distributions) based on overlap of ranked peaks (consistency)

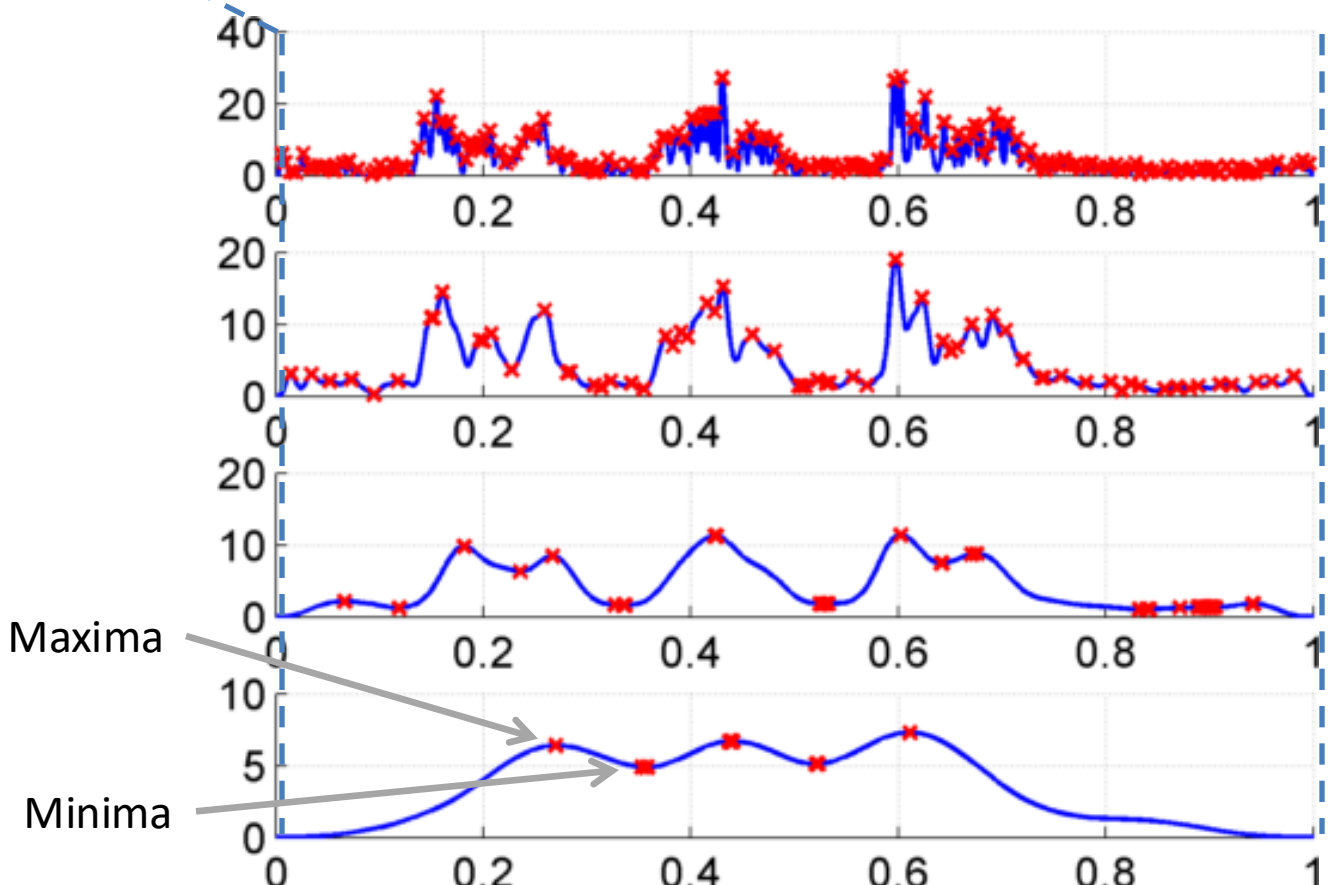




Genome Research 2012, 22:1813-1831

Multiscale Analysis, Minima/Maxima based Coarse Segmentation





Window Length

1kb

4kb

16kb

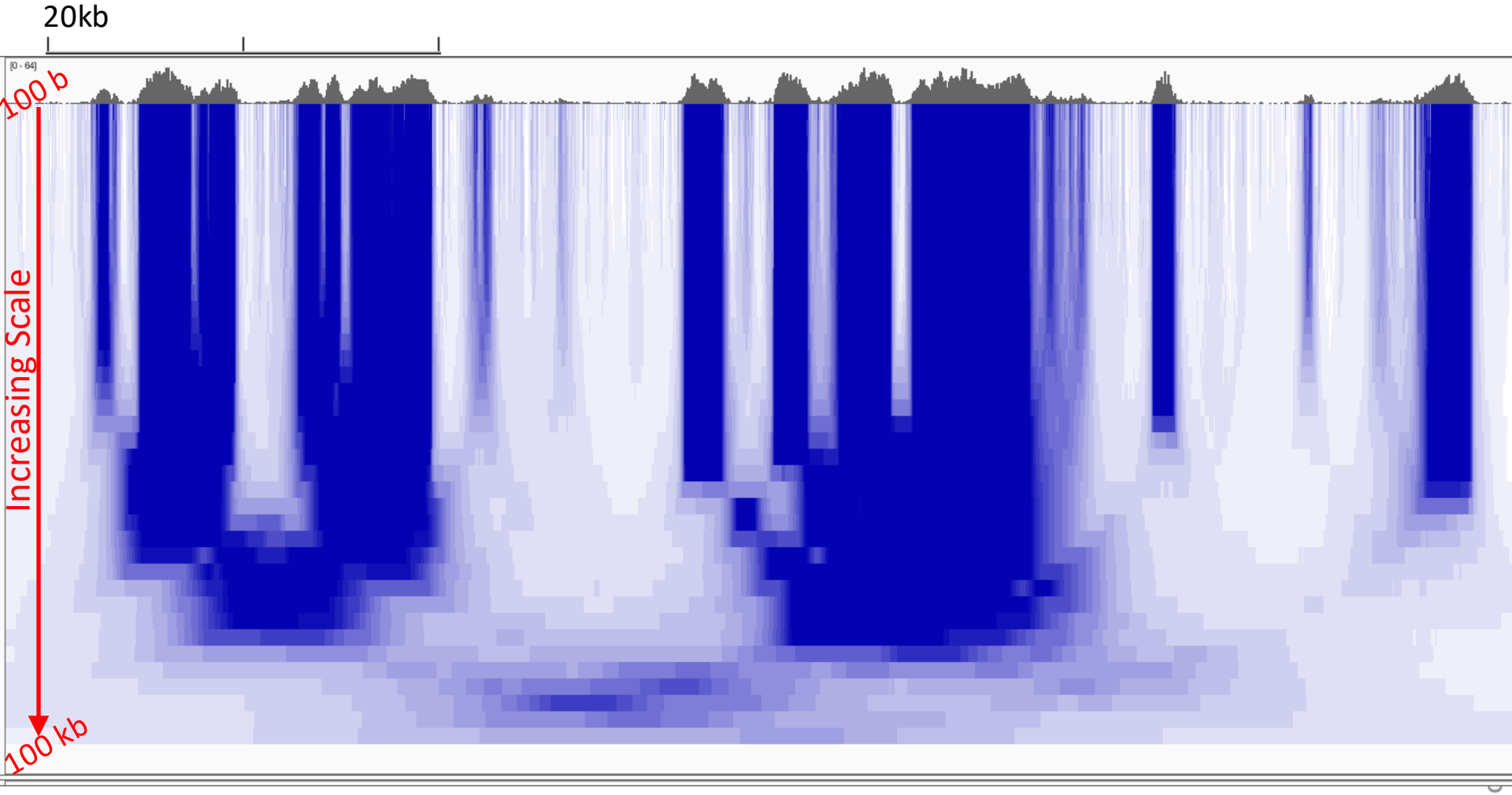
64kb

Maxima

Minima

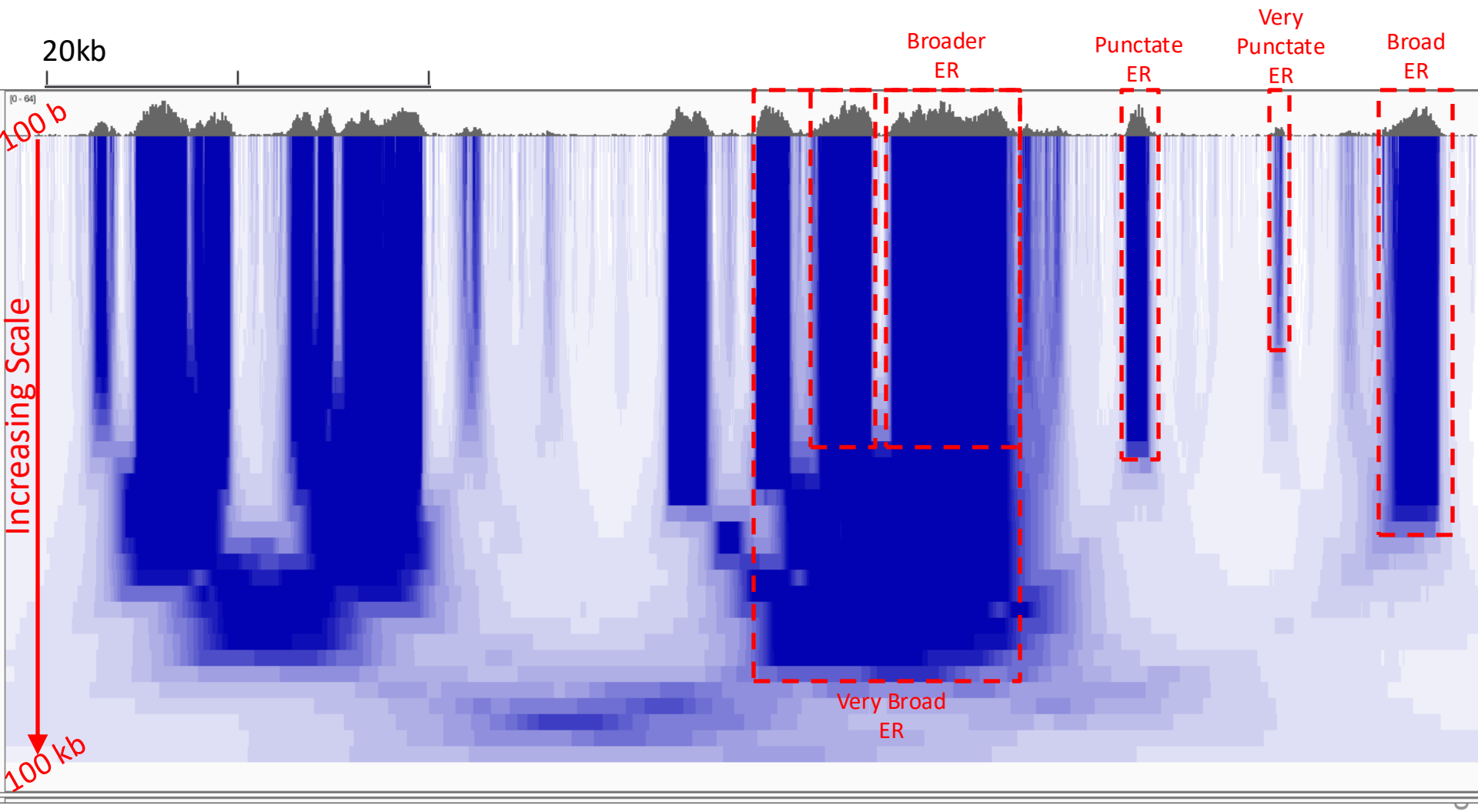
Harmanci et al, Genome Biology 2014, MUSIC.gersteinlab.org

Multiscale Decomposition



[Harmanci *et al, Genome Biol.* ('14)]

Multiscale Decomposition



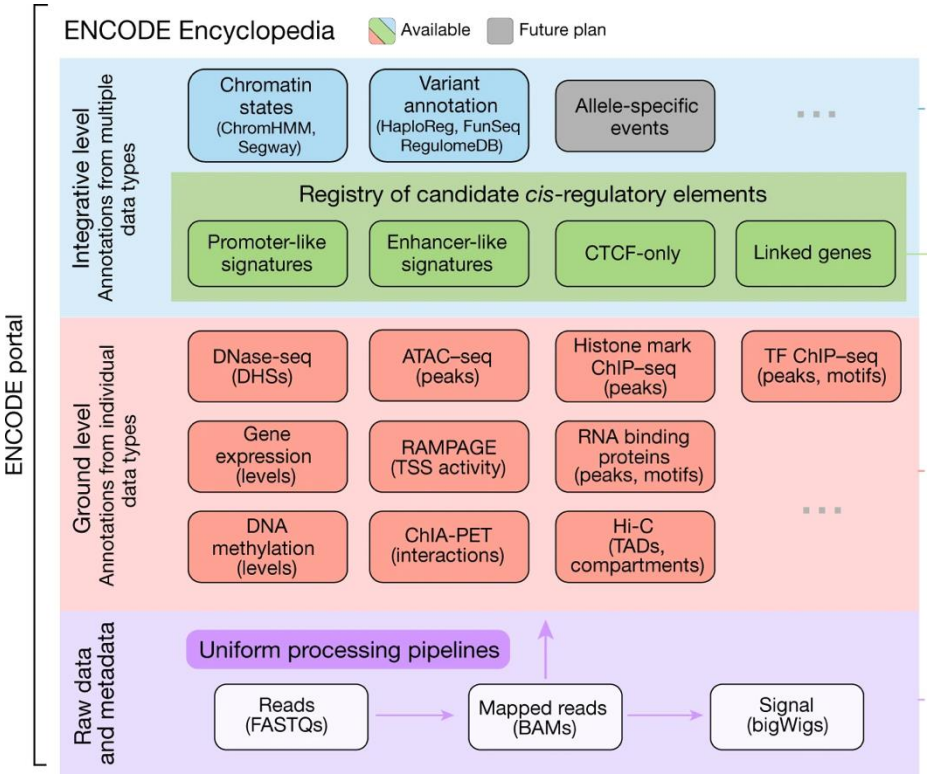
[Harmanci *et al, Genome Biol.* ('14)]

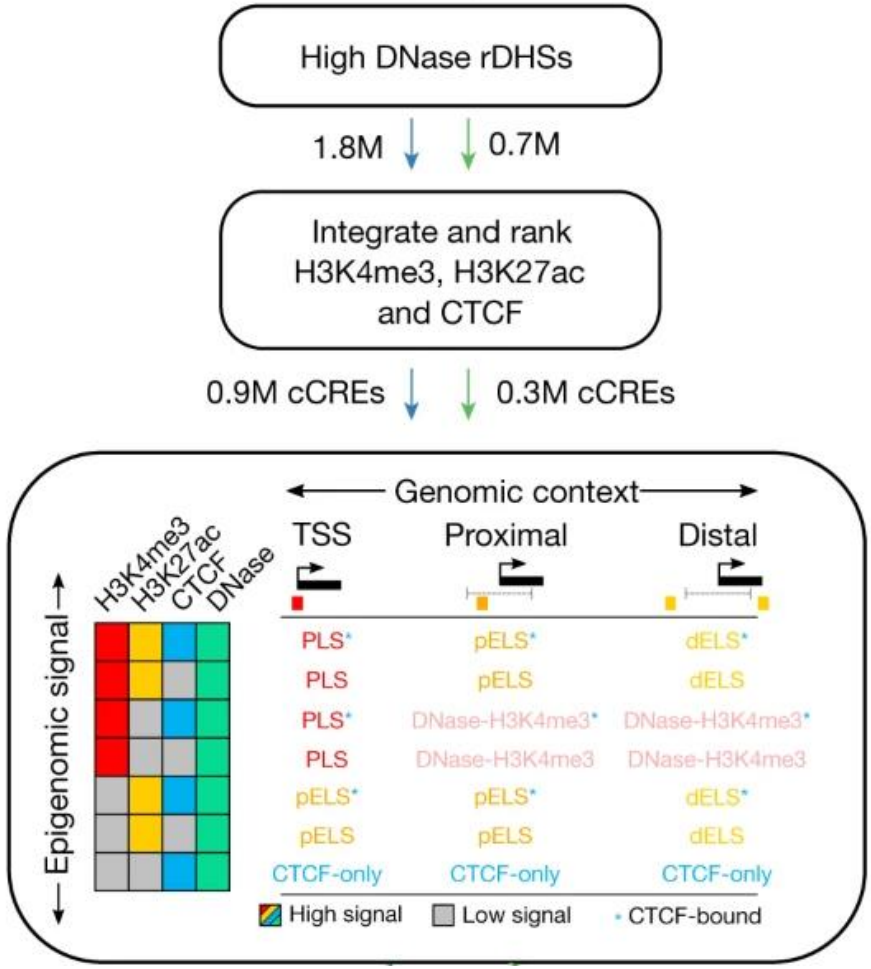
Outline

- **Part 1 : Generic Annotation (not related to an individual's variants)**
  - RNA-seq, Chip-seq
  - Integration
  - , Hi-C
- Part 2 : Annotation related to an individual's variants
  - ASE/ASB
  - GWAS & eQTL

Simple Integration for Elements

Broad ENCODE Annotation





[Encode Consortium et al. Nature ('20)]

Background on computational annotation for non-coding regions

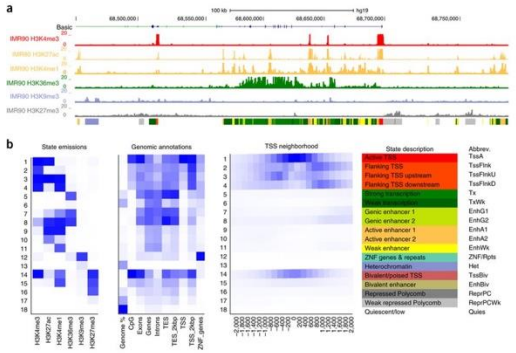
- **Peak calling**:
  - ✓PeakSeq, SPP, MACS2, Hotspot …
  - ✓ENCODE Encyclopedia

- **Genome segmentation**: partition the genome into regions (states) with distinct epigenomic profiles, then assign each state a functional label.
  - ✓ChromHMM: Multivariate Hidden Markov Model
  - ✓Segway: Dynamic Bayesian Network Model

- Supervised regulatory prediction: learn predictive models from labeled dataset of regulatory elements.
  - ✓ CSI-ANN: Time-Delay Neural Network
  - ✓ RFECS: Random Forest
  - ✓ DEEP: Ensemble SVM + Artificial Neural Network
  - ✓ REPTILE: Random Forest
  - ✓ gkm-SVM: Gapped k-mer
  - ✓Matched-Filter: Signal Processing Filters

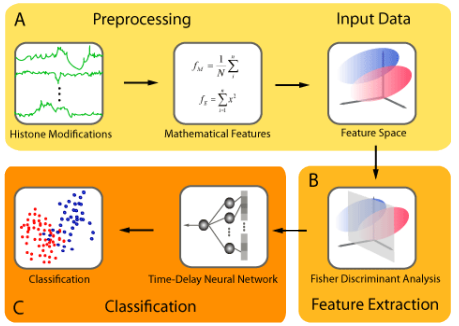
- **Target finding**
  - ✓ Ripple, TargetFinder, JEME, PreSTIGE, IM-PET

ChromHMM



J. Ernst, M. Kellis. *Nat. Protoc., 2017*

CSI-ANN



H.A. Firpi, D. Ucar, K. Tian. Bioinformatics, 2010

Outline

- **Part 1 : Generic Annotation (not related to an individual's variants)**
  - RNA-seq, Chip-seq
  - Integration
  - , Hi-C
- Part 2 : Annotation related to an individual's variants
  - ASE/ASB
  - GWAS & eQTL

Hi-C

3D organization of genome

image credit: Iyer et al. BMC Biophysics 2011, cartoonist John Chase

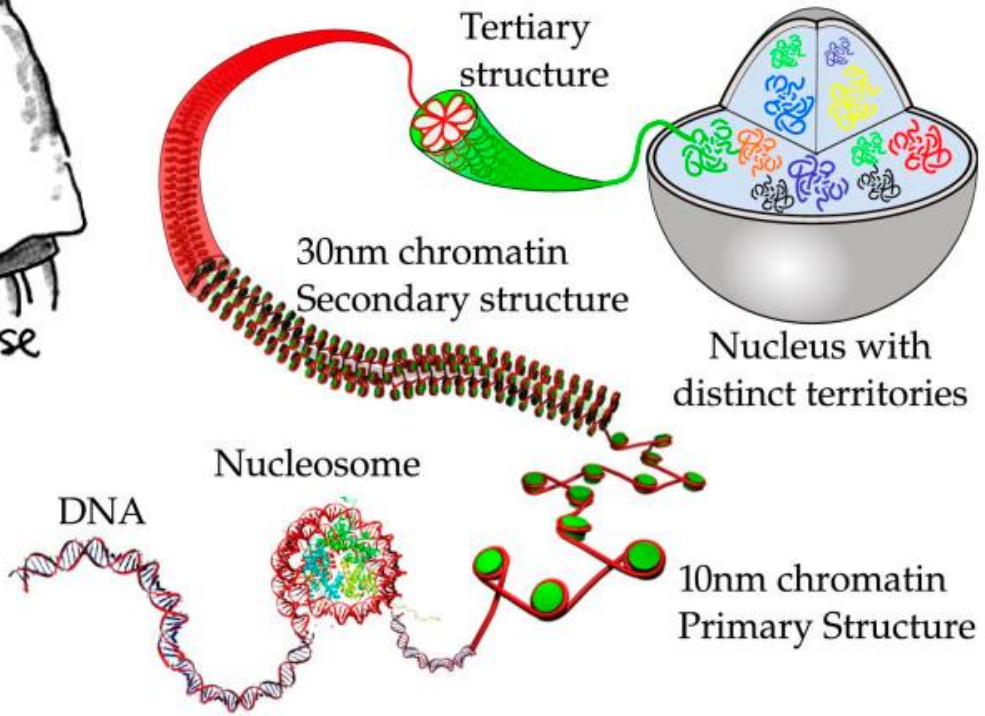
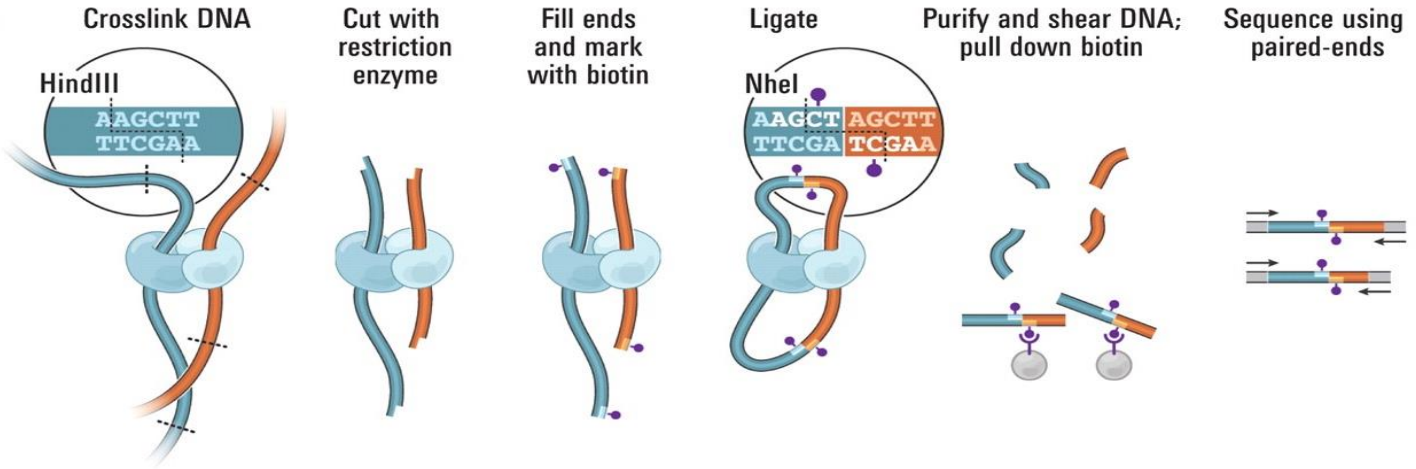
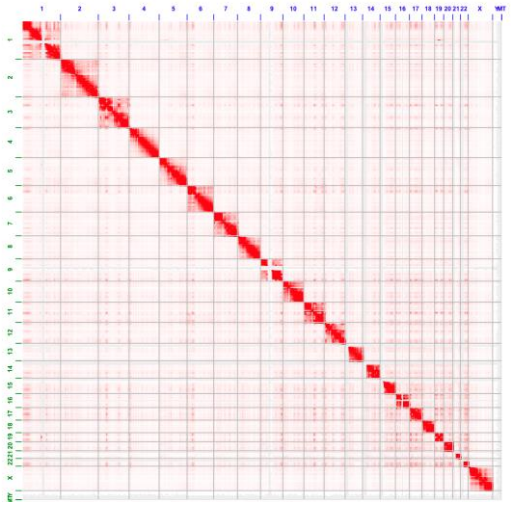


image credit: Iyer et al. BMC Biophysics 2011

Hi-C contact map

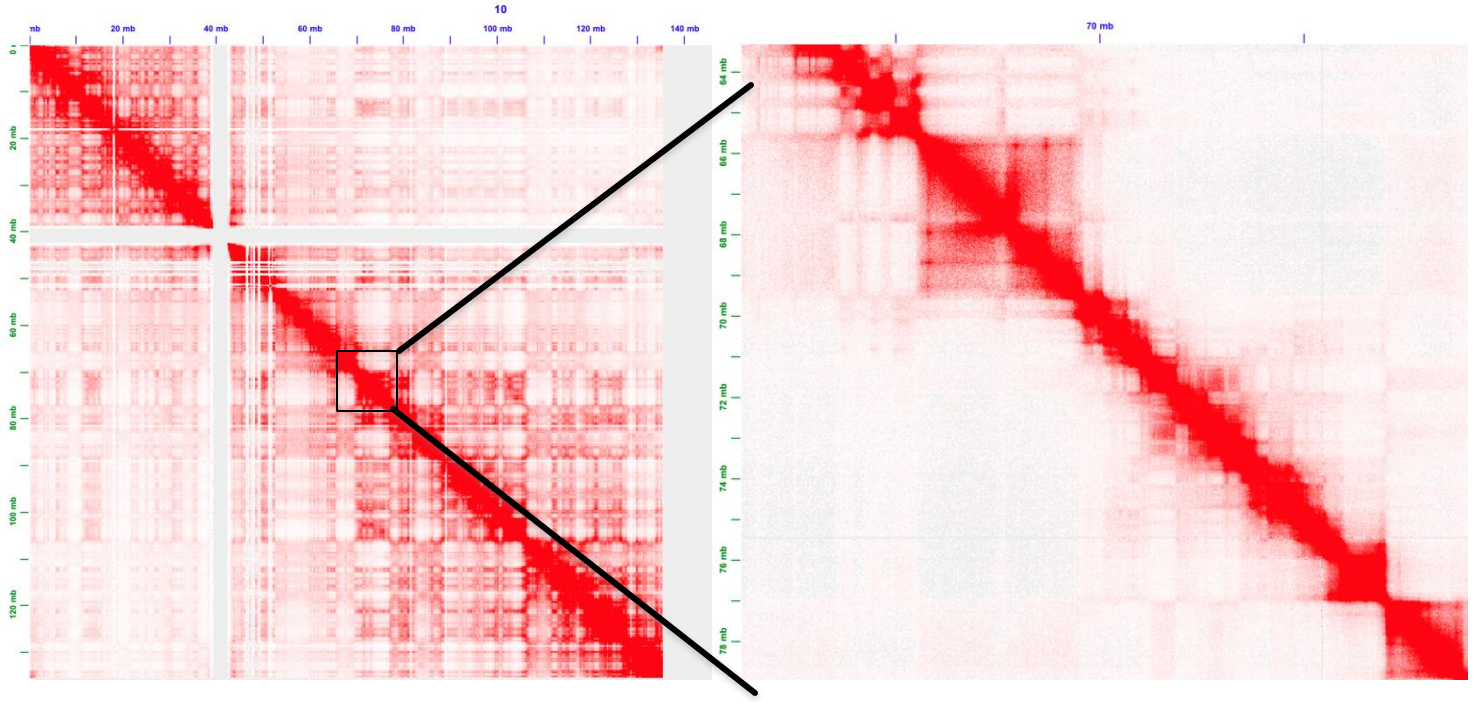




A network can be expressed as an adjacency matrix **A** where $\mathbf{A}_{ij} \neq 0$ if *i* and *j* are in contact & 0 otherwise

Science 2009, 5950: 289-293

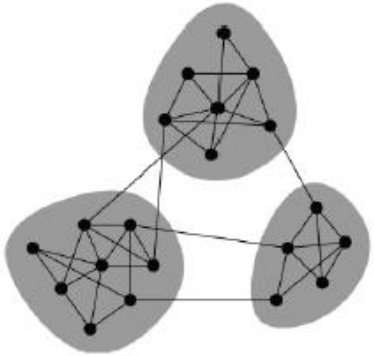
Topologically associating domains (TADs)



TADs have apparent hierarchical organization

(More Later)

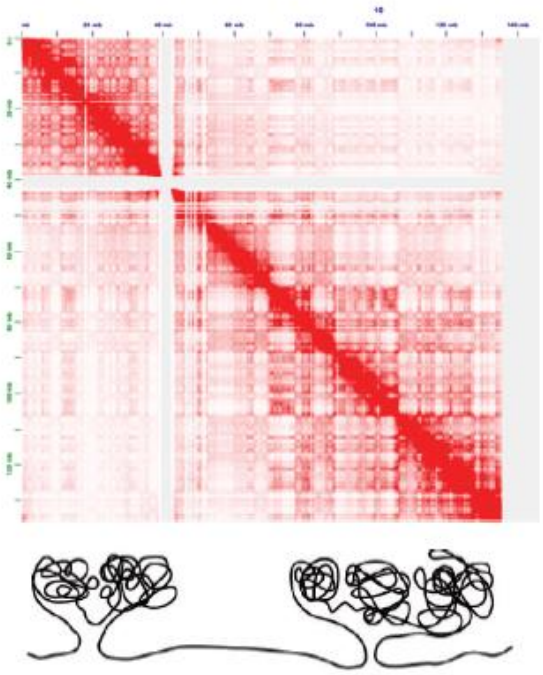
Identifying TADs in multiple resolutions



Modularity maximization

$$Q = \frac{1}{2m} \sum_{i,j} \left(W_{ij} - \frac{k_i k_j}{2m} \right) \delta_{\sigma_i \sigma_j}$$

network	contact map
node	chromosome bin
edge	Hi-C contact
# of connections	coverage
module	domain



schematic adapted from ref. [2]

To be continued in network section…..

[Yan et al., *PLOS Comp. Bio.* (in revision, ‘17); bioRxiv 097345]

References for 25m7 part-1 (Functional Annotation of the Genome)

- Alexander, R. P., Fang, G., Rozowsky, J., Snyder, M., & Gerstein, M. B. (2010).
  Annotating non-coding regions of the genome.
  Nature Reviews Genetics, 11(8), 559–571.
  https://doi.org/10.1038/nrg2814
  (Read whole thing.)

- Expanded encyclopaedias of DNA elements in the human and mouse genomes
  The ENCODE Project Consortium et al. Nature volume 583, pages 699–710 (2020)
  https://www.nature.com/articles/s41586-020-2493-4
  (Focus on text associated with Figs 2 and 3.)

- Mackenzie, R. (2024, January 24).
  RNA-Seq: Basics, applications and Protocol.
  Genomics Research From Technology Networks.
  https://www.technologynetworks.com/genomics/articles/rna-seq-basics-applications-and-protocol-299461
  (Extra reference, optional)