Harnessing AI to Make Sense of Large, Complex Datasets & Re-Invigorate R&D Innovation:

Evolution of Element Annotation, from Calling ChIP Peaks to Determining Genome Folding

Mark Gerstein

Yale

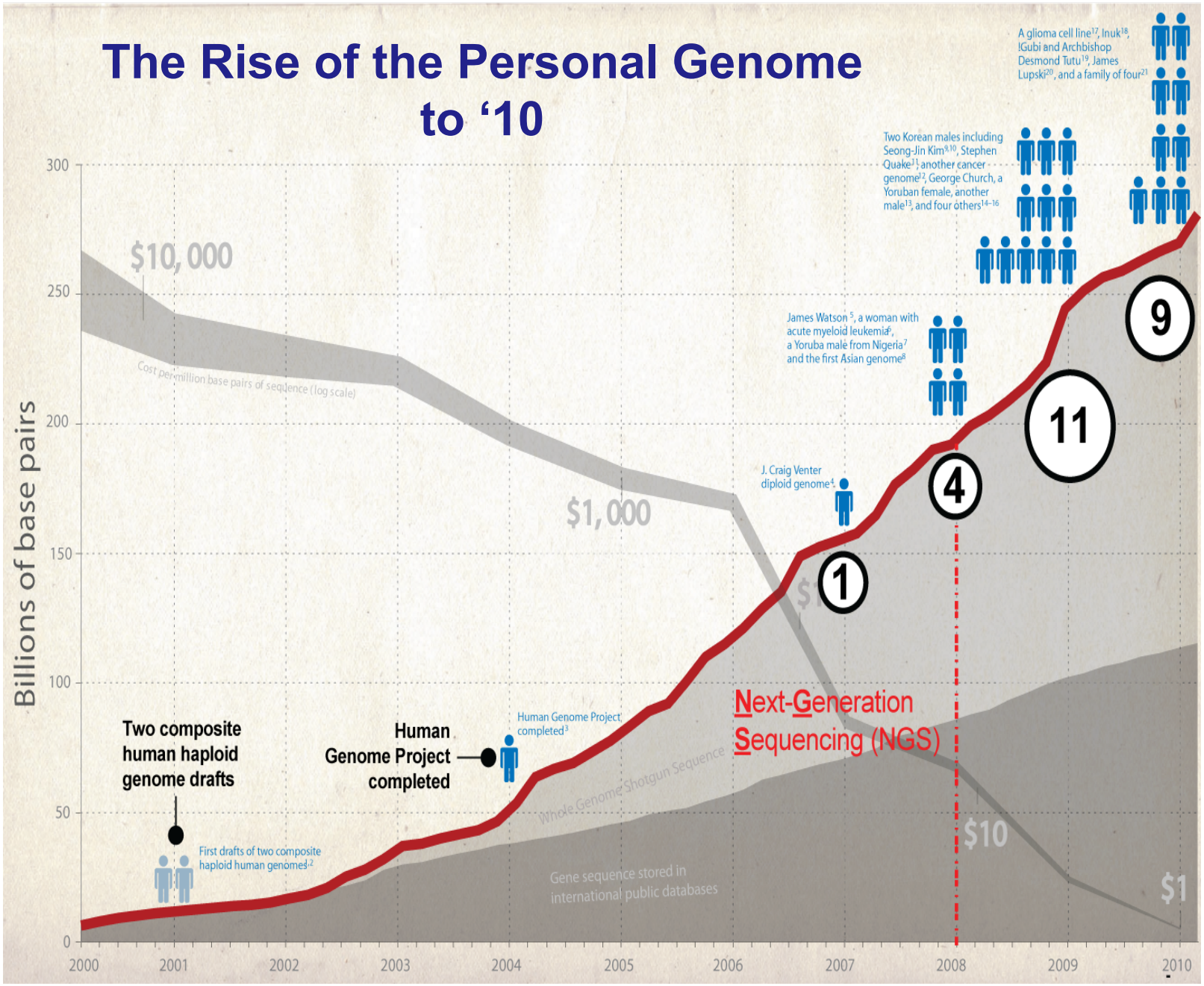
Slides "tweetable" (via @markgerstein). See last slide for more info & freely downloadable from **Lectures.GersteinLab.org**

High Level Ideas on the Evolution of Machine Learning in Genome Analysis

- Provides an illustration of how machine learning functions to make sense of large, complex datasets

- The problem of annotating active & repressed regions in the genome
  - Original formulation in terms of “peak calling” on the linear genome
  - Revision of the original work, now at multi-scale
  - Recent radical change: now thinking of the genome as a 3D folded molecule

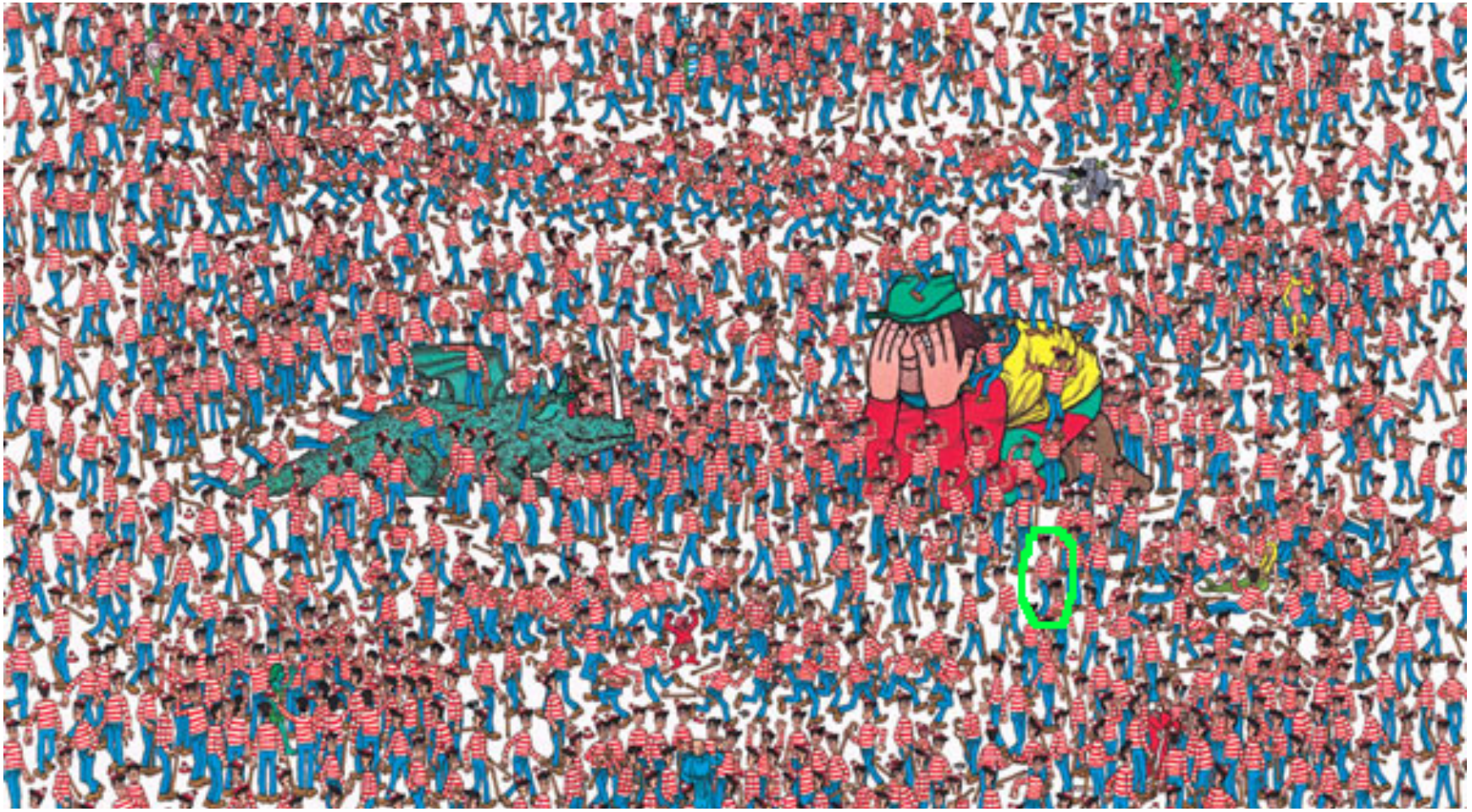
The Rise of the Personal Genome to '10



Adapted from *Nature* 2010

Where is Waldo?

(Finding the key mutations in ~3M Germline variants & ~5K Somatic Variants in a Tumor Sample)



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

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.

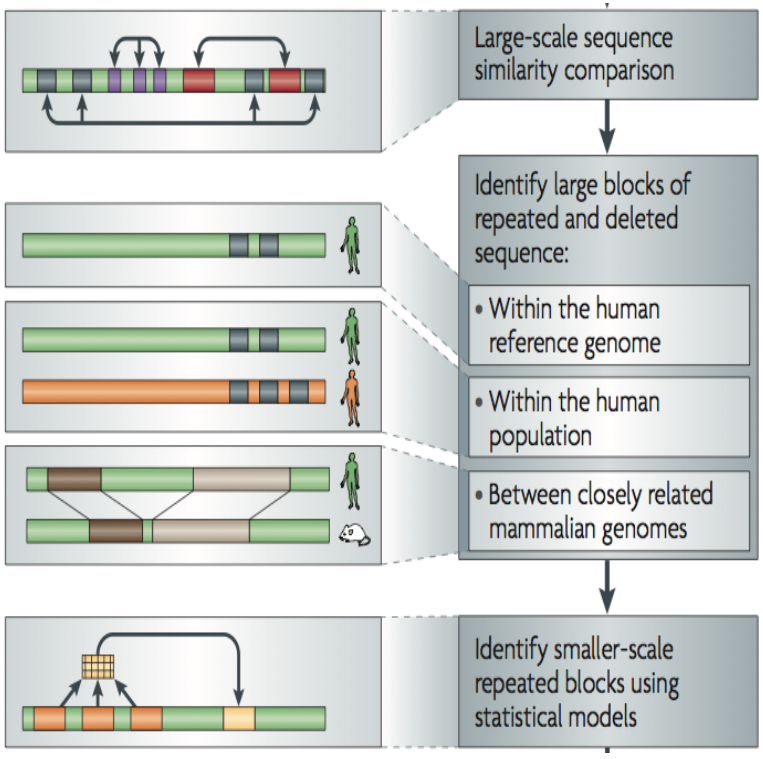
- There appear to be about 30,000–40,000 protein-coding genes in the human genome—only about twice as many as in worm or fly. However, the genes are more complex, with more alternative splicing generating a larger number of protein products.
- The full set of proteins (the 'proteome') encoded by the 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

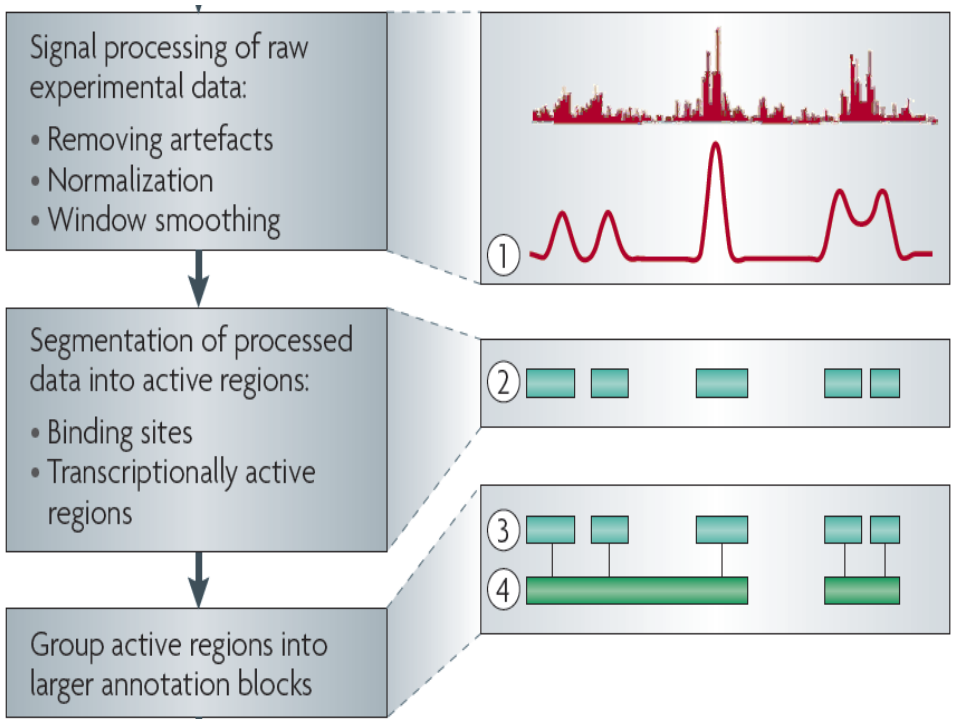
Non-coding Annotations: Overview

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

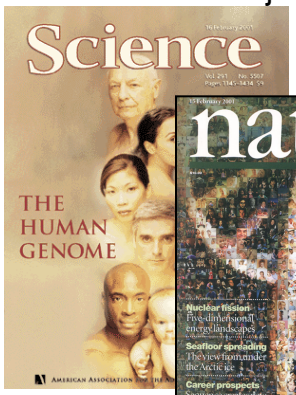
Sequence features, incl. **Conservation**

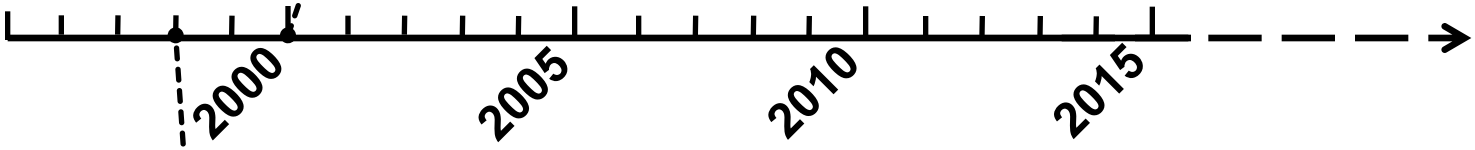
Functional Genomics
Chip-seq (Epigenome & seq. specific TF)
and ncRNA & un-annotated transcription





The Human Genome Project





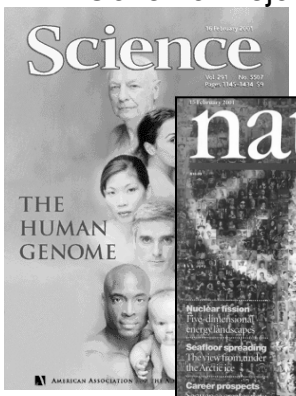
Worm Genome

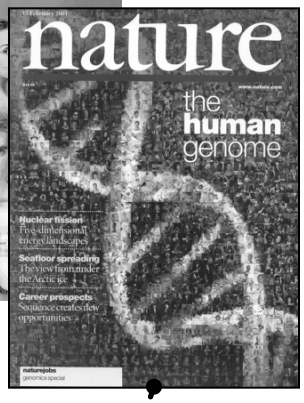
The Human Genome Project

ENCODE Pilot

ENCODE Production

Comparative ENCODE



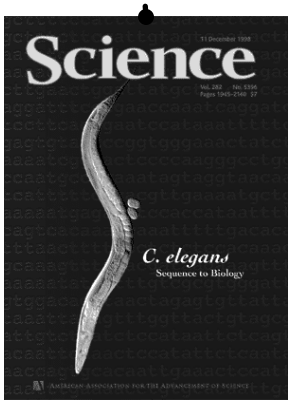


2000

2005

2010

2015



Worm Genome

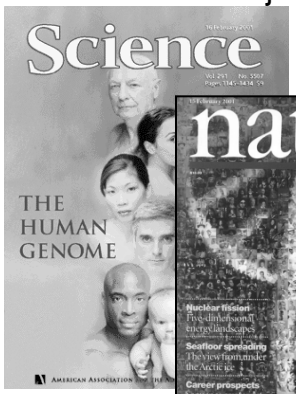
modENCODE

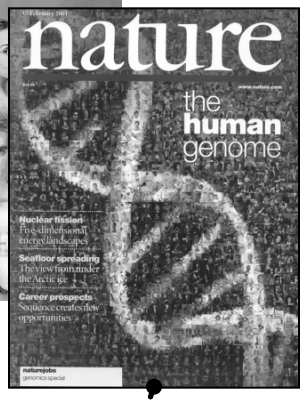
The Human Genome Project

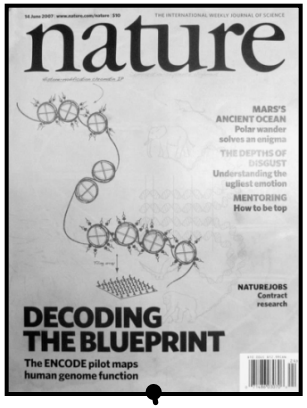
ENCODE Pilot

ENCODE Production

Comparative ENCODE





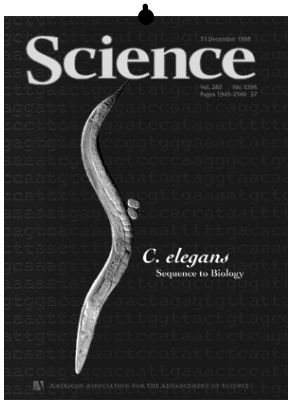


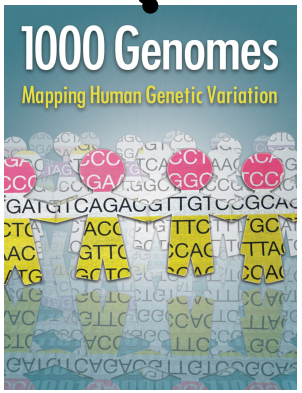
2000

2005

2010

2015





Worm Genome

modENCODE

1000 Genomes Pilot

1000 Genomes Production

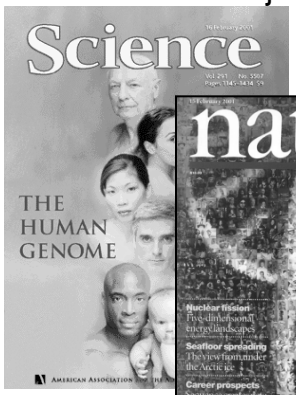
The Human Genome Project

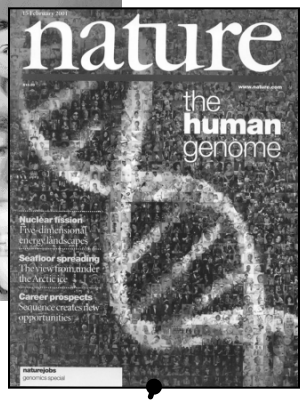
ENCODE Pilot

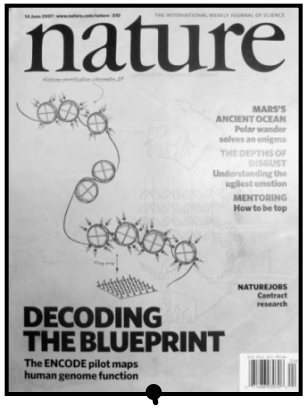
ENCODE Production

Comparative ENCODE

Epigenome Roadmap





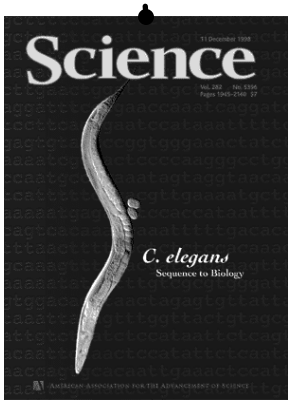


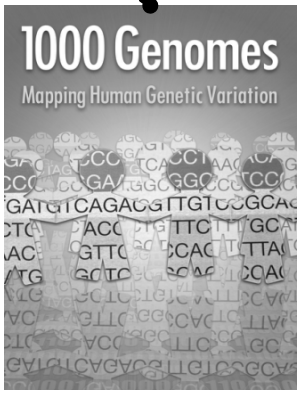
2000

2005

2010

2015





Worm Genome

modENCODE

1000 Genomes Pilot

1000 Genomes Production

GTEx

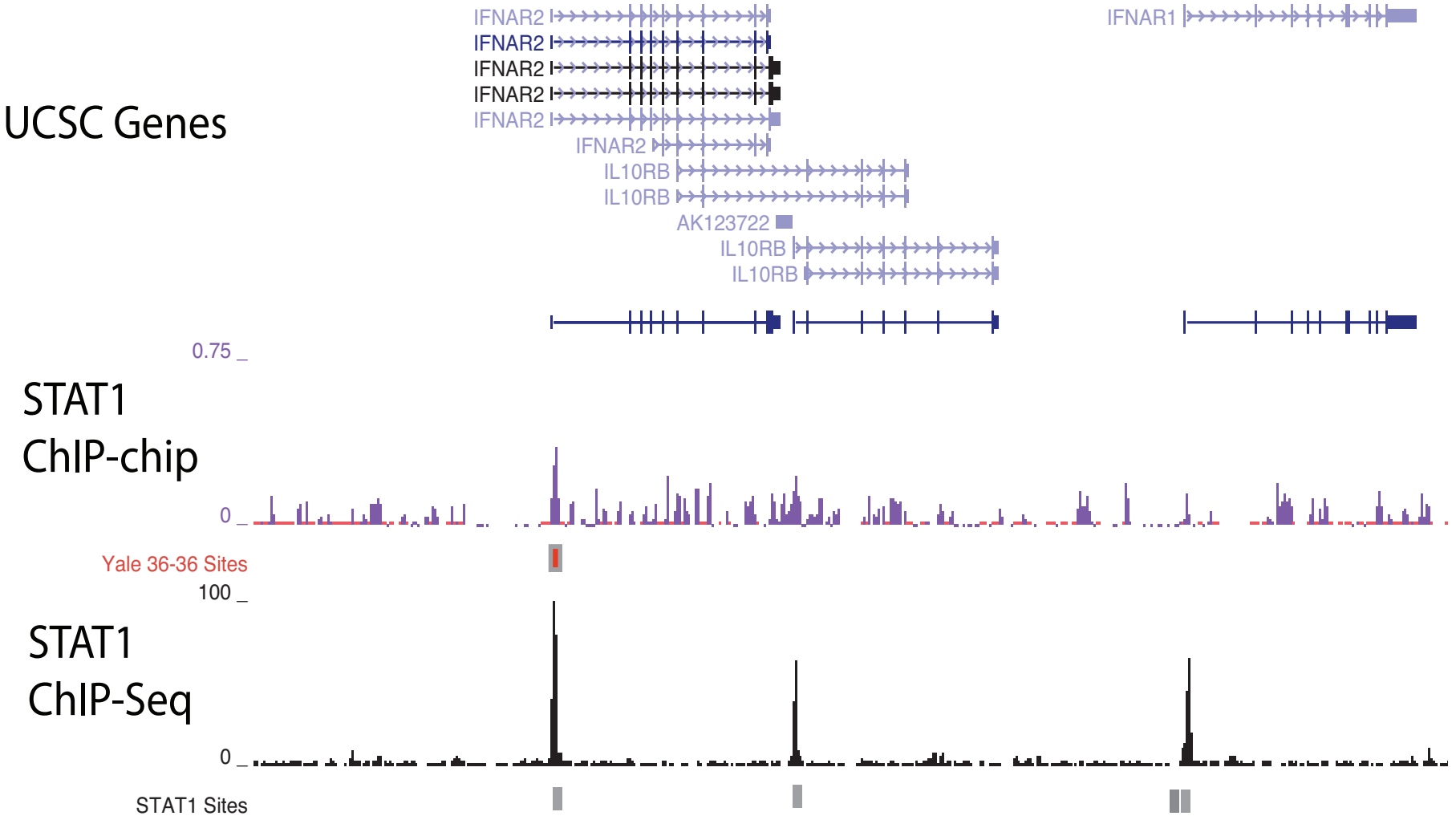
Evolution of Element Annotation,
from Calling ChIP Peaks to Determining Genome Folding

- Characterizing Regulatory Sites on the Linear Genome
  - Original peak calling approach (with PeakSeq)
  - New Multi-scale "site" calling (with Music)
- Characterizing TADs from 3D Genome Folding
  - Using modularity for identification, at multiple scales (with MrTADFinder)
  - Developing an appropriate null expectation
- Features of Multi-resolution TADs
  - Specific TFs & HMs associated with TAD boundaries at different scales
  - Assoc. strong enough to build a predictor
  - HOT regions at boundaries
  - Relation to somatic mutations
- Technical Analysis of TADs
  - Spectral analysis quantifying reproducibility of Hi-C data sets (with HiC-Spector)

Evolution of Element Annotation, from Calling ChIP Peaks to Determining Genome Folding

- Characterizing Regulatory Sites on the Linear Genome
  - Original peak calling approach (with PeakSeq)
  - New Multi-scale "site" calling (with Music)
- Characterizing TADs from 3D Genome Folding
  - Using modularity for identification, at multiple scales (with MrTADFinder)
  - Developing an appropriate null expectation
- Features of Multi-resolution TADs
  - Specific TFs & HMs associated with TAD boundaries at different scales
  - Assoc. strong enough to build a predictor
  - HOT regions at boundaries
  - Relation to somatic mutations
- Technical Analysis of TADs
  - Spectral analysis quantifying reproducibility of Hi-C data sets (with HiC-Spector)

ChIP-seq vs ChIP-chip: Much cleaner signal from sequencing than arrays

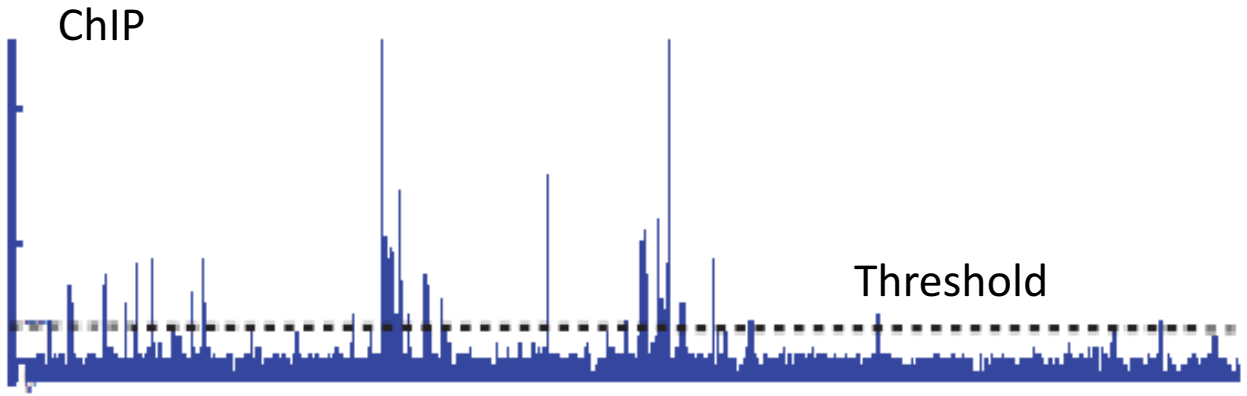


[Rozowsky et al. Nat. Biotech ('09)]

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)



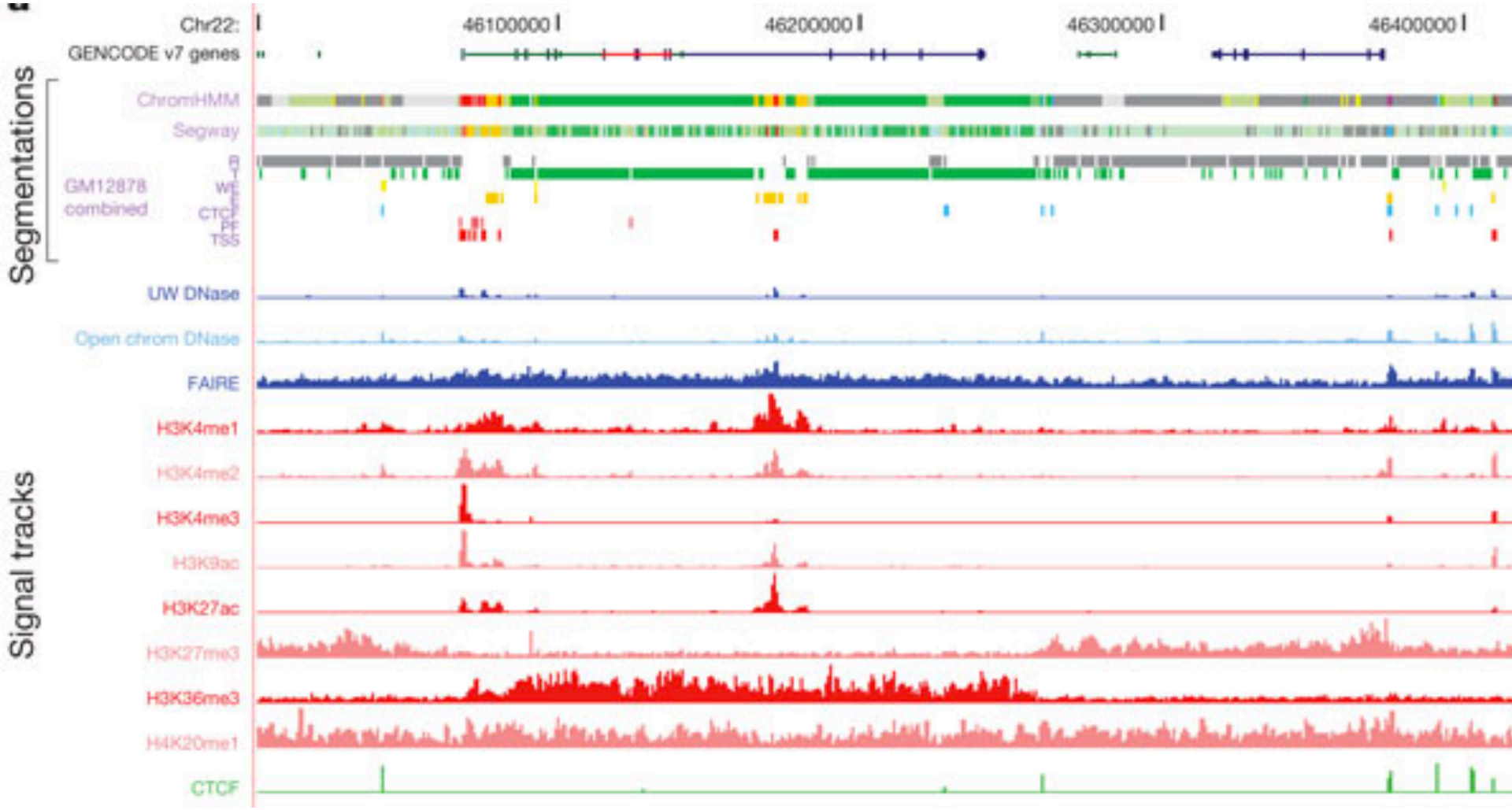
Potential Targets →

Score against the **control**

Significantly Enriched targets →

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

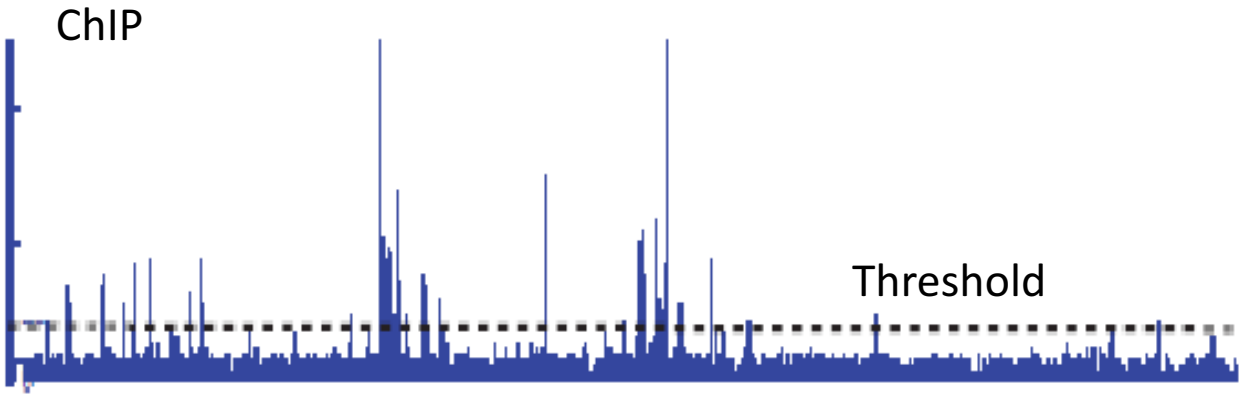
Multi-track analysis: Segmentation



[Encode Consortium (’12), Nature; Ernst & Kellis, Hoffman & Noble]

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 →

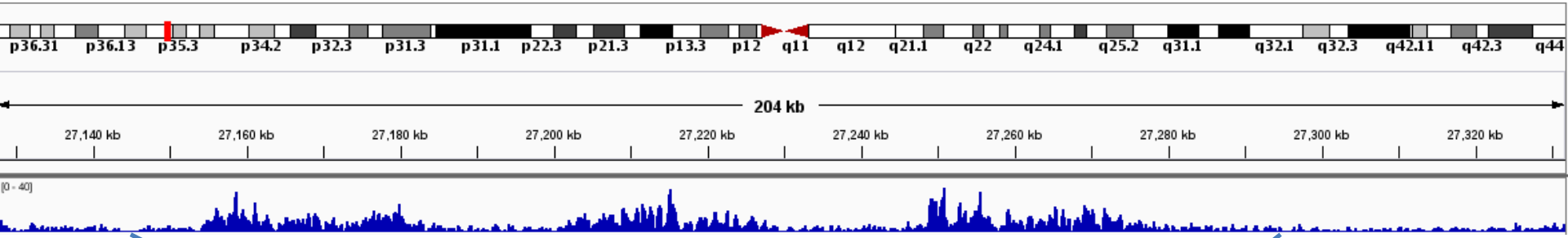
- Score against the control

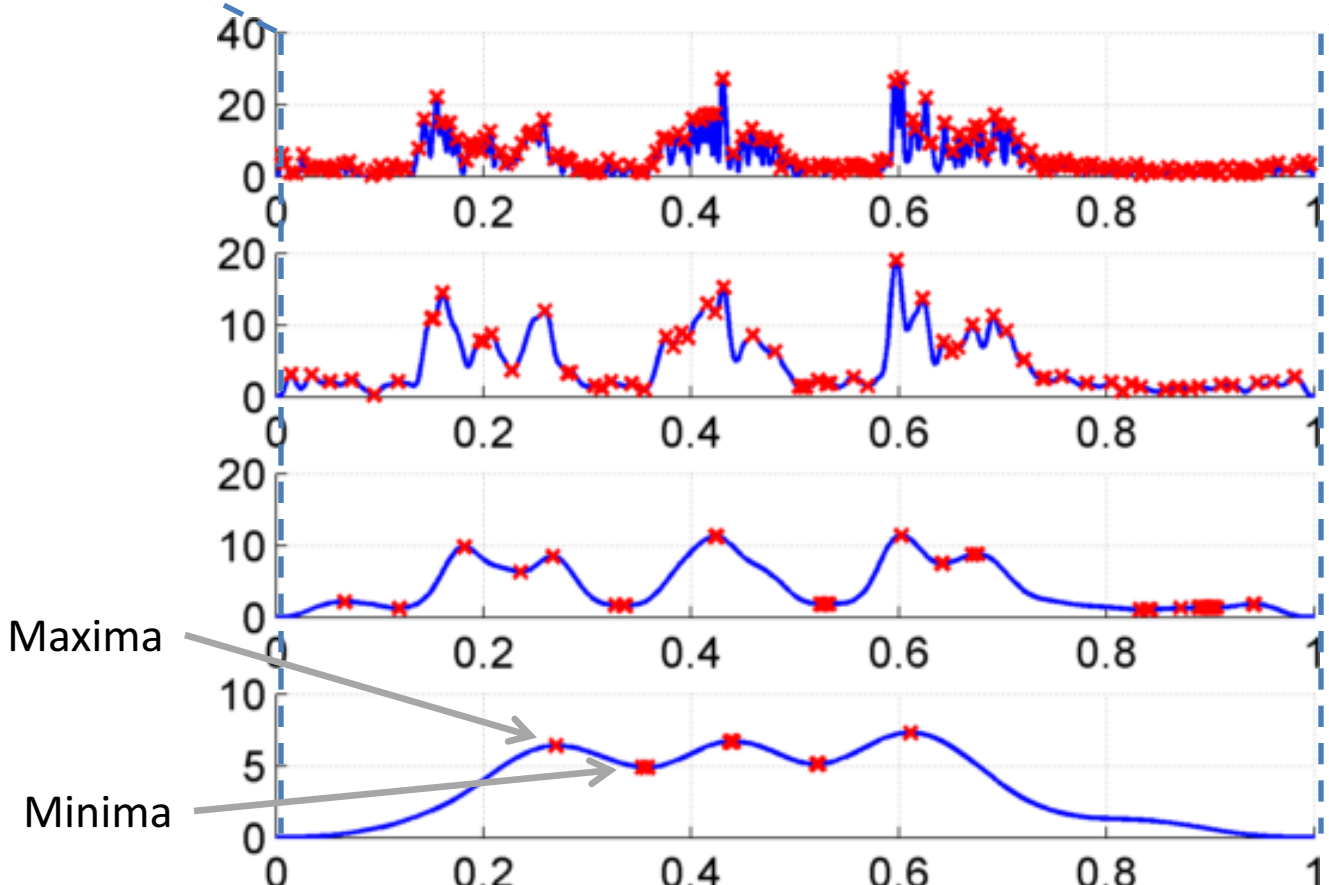
Significantly Enriched targets →

Now an update: "PeakSeq 2" => MUSIC

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

Multiscale Analysis, Minima/Maxima based Coarse Segmentation





Window Length

1kb

4kb

16kb

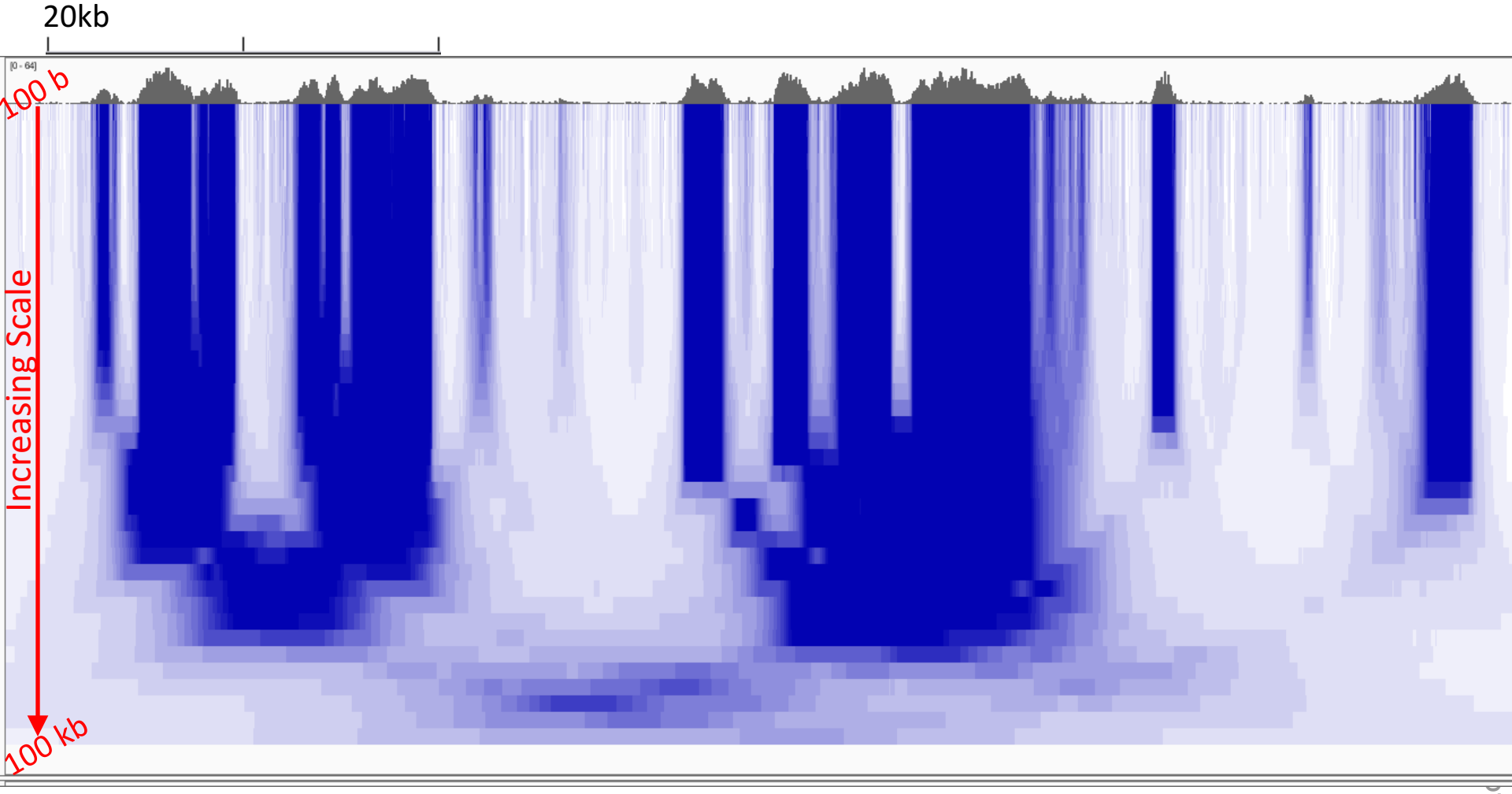
64kb

Maxima

Minima

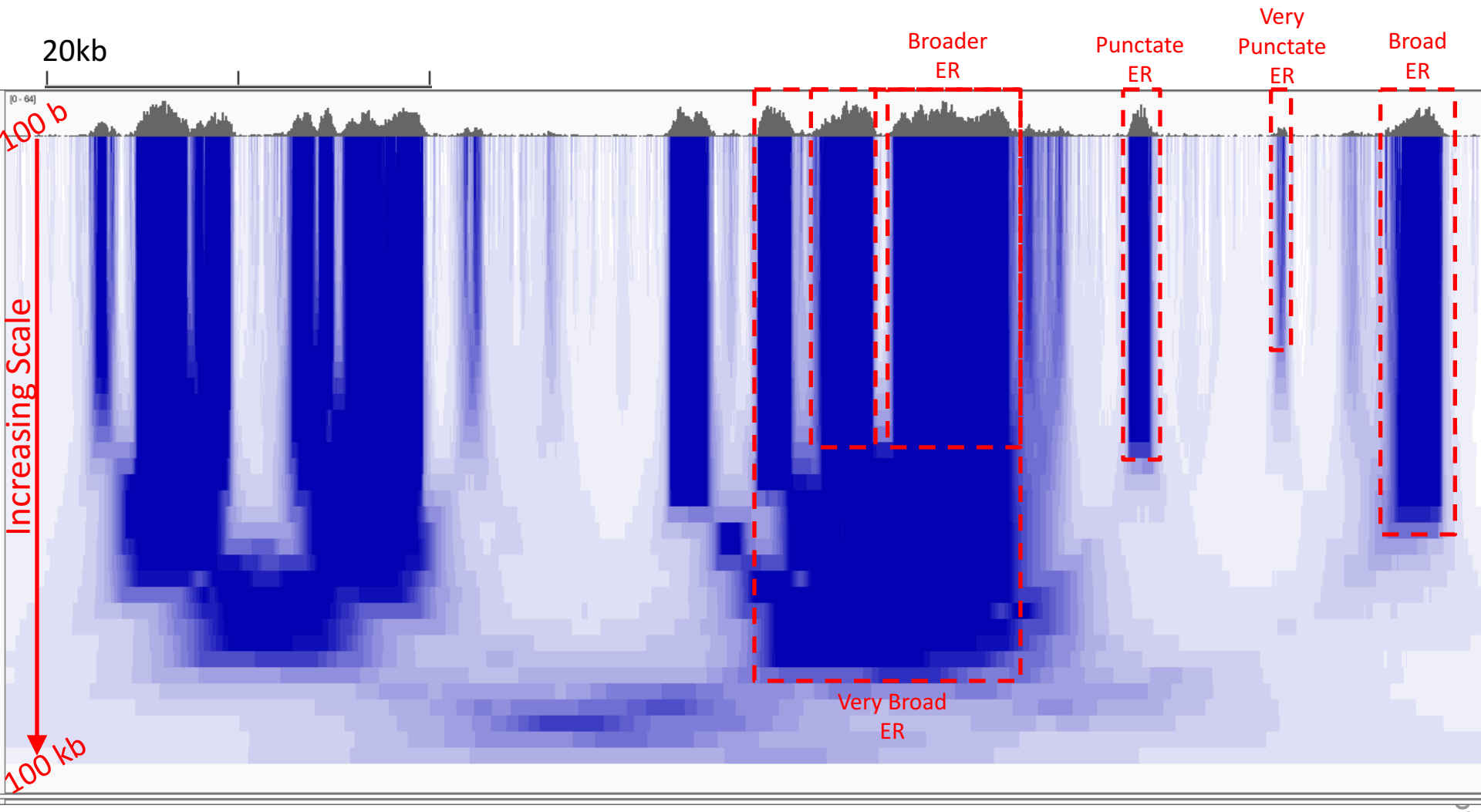
Harmanci et al, Genome Biology 2014, MUSIC.gersteinlab.org

Multiscale Decomposition



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

Multiscale Decomposition



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

Evolution of Element Annotation, from Calling ChIP Peaks to Determining Genome Folding

- Characterizing Regulatory Sites on the Linear Genome
  - Original peak calling approach (with PeakSeq)
  - New Multi-scale "site" calling (with Music)
- Characterizing TADs from 3D Genome Folding
  - Using modularity for identification, at multiple scales (with MrTADFinder)
  - Developing an appropriate null expectation
- Features of Multi-resolution TADs
  - Specific TFs & HMs associated with TAD boundaries at different scales
  - Assoc. strong enough to build a predictor
  - HOT regions at boundaries
  - Relation to somatic mutations
- Technical Analysis of TADs
  - Spectral analysis quantifying reproducibility of Hi-C data sets (with HiC-Spector)

3D organization of genome

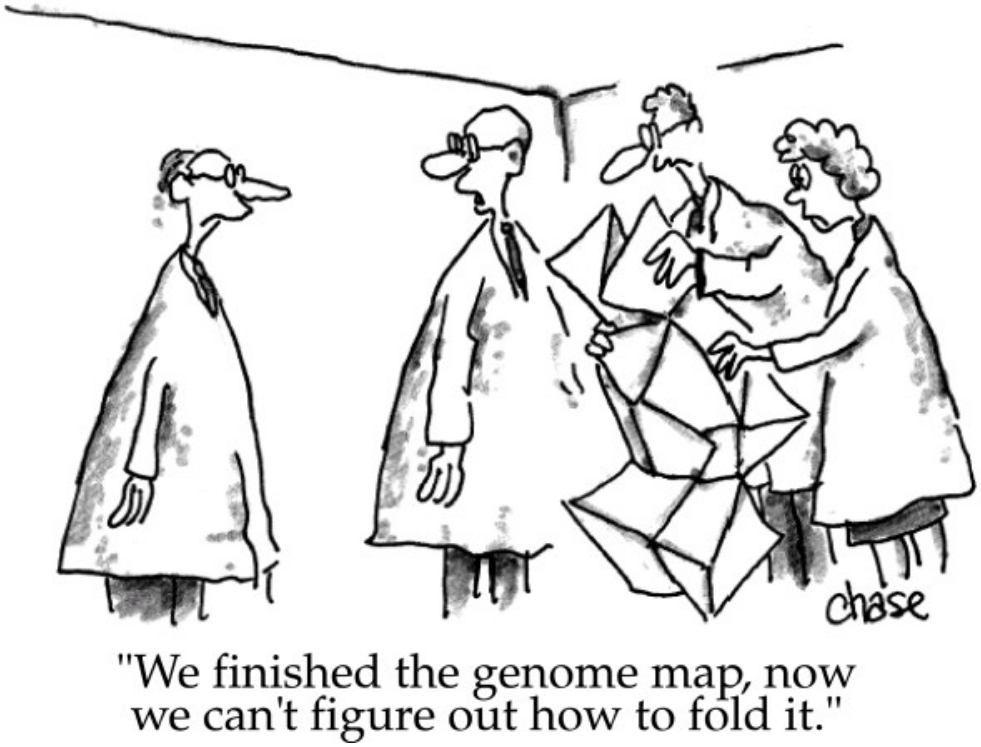


image credit: Iyer et al. BMC Biophysics 2011, cartoonist John Chase

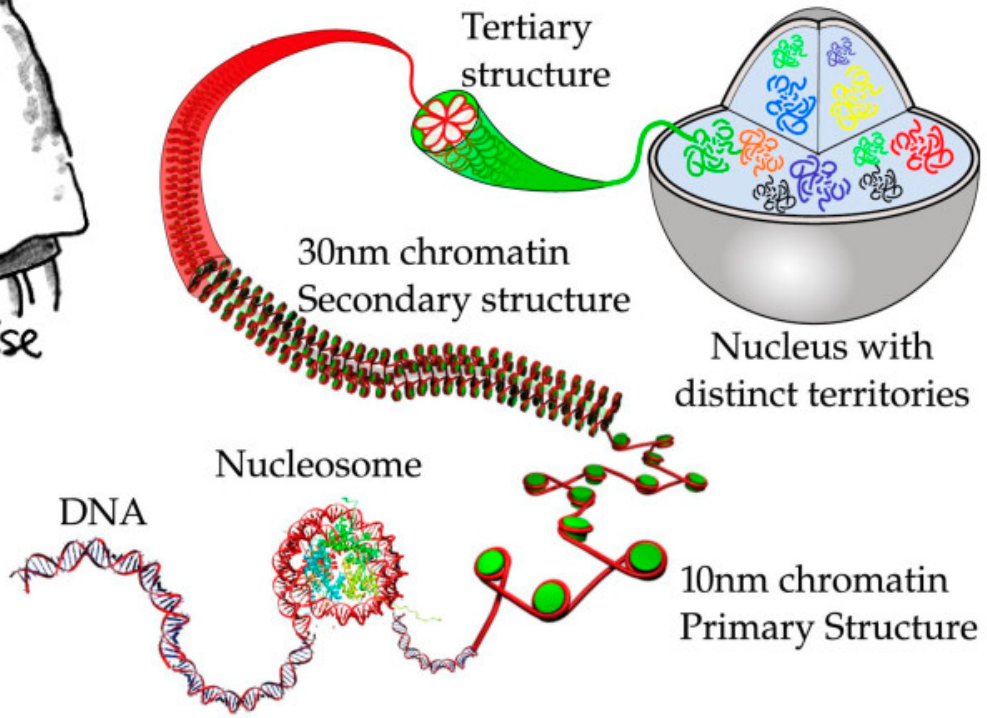


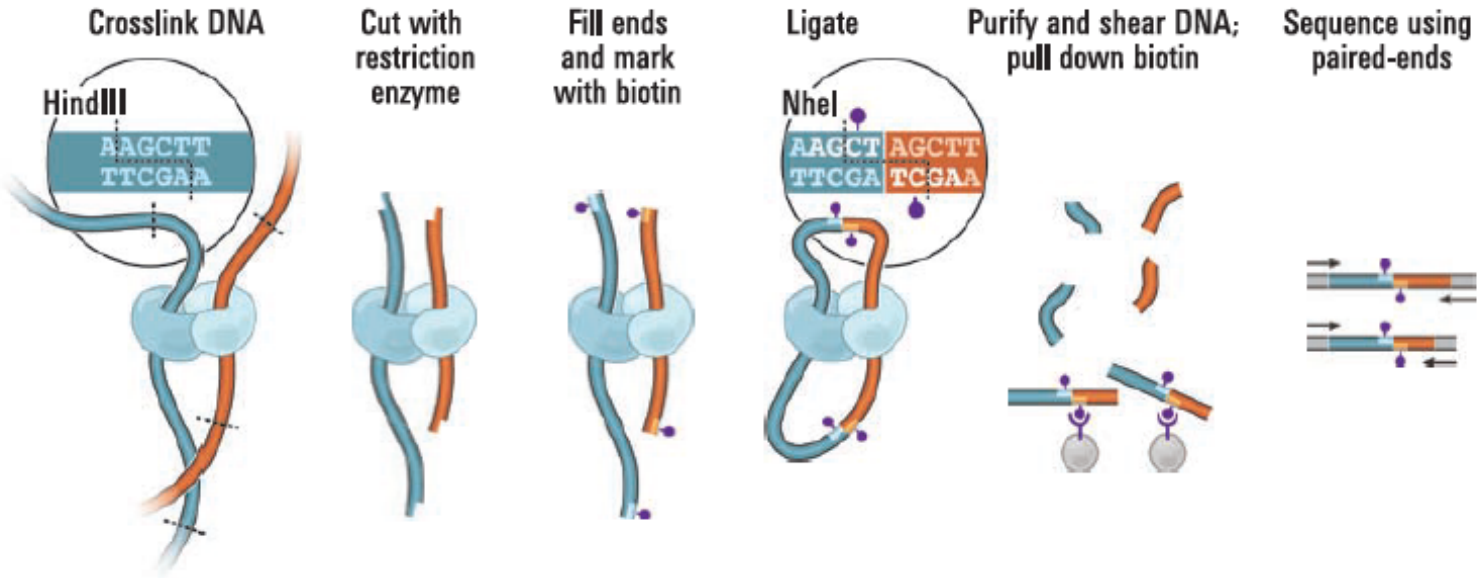
image credit: Iyer et al. BMC Biophysics 2011

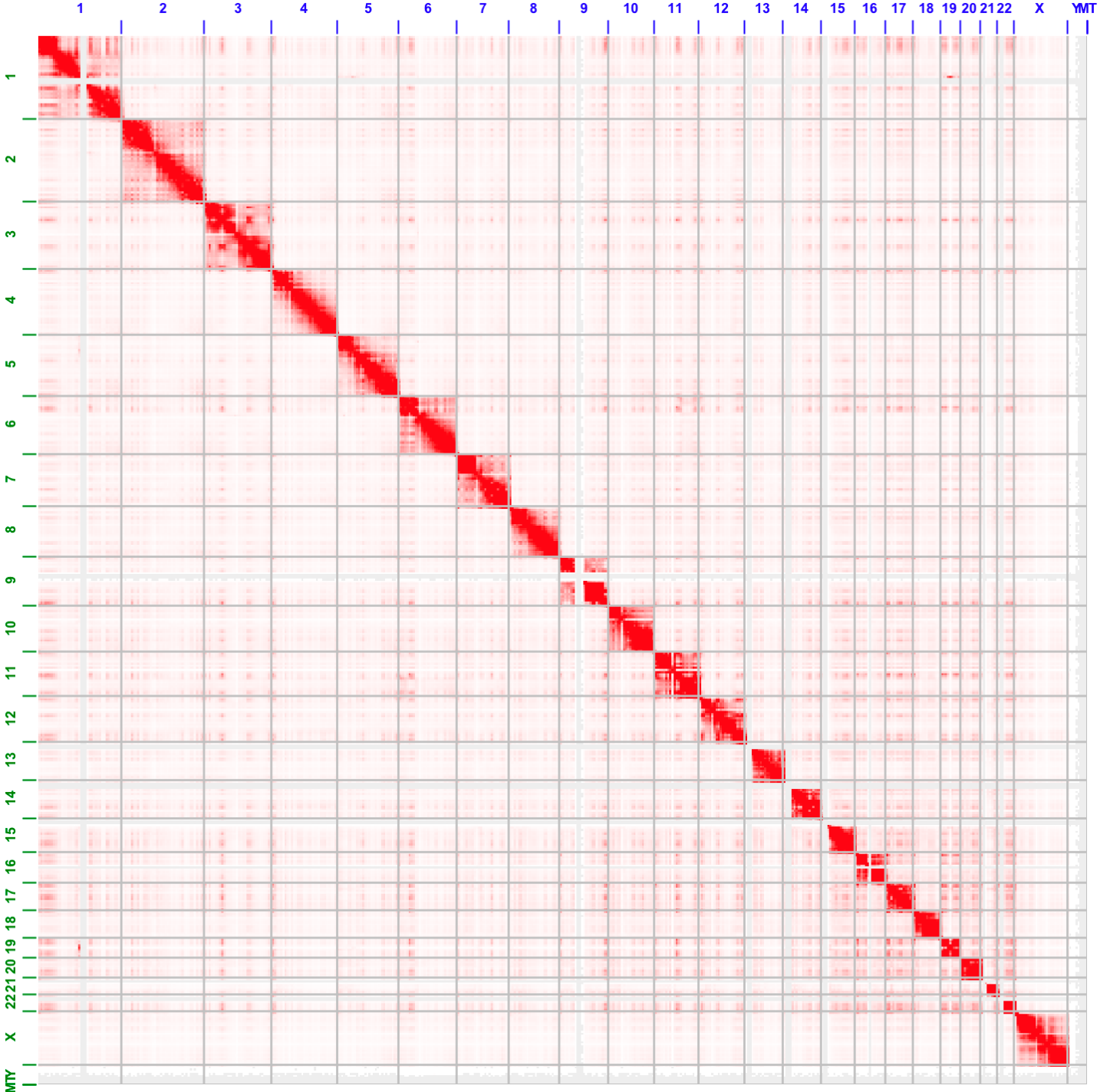
Chromosome conformation capture (3C) and Hi-C

Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome

Erez Lieberman-Aiden,[1,2,3,4]* Nynke L. van Berkum,[5]* Louise Williams,[1] Maxim Imakaev,[2] Tobias Ragoczy,[6,7] Agnes Telling,[6,7] Ido Amit,[1] Bryan R. Lajoie,[5] Peter J. Sabo,[8] Michael O. Dorschner,[8] Richard Sandstrom,[8] Bradley Bernstein,[1,9] M. A. Bender,[10] Mark Groudine,[6,7] Andreas Gnirke,[1] John Stamatoyannopoulos,[8] Leonid A. Mirny,[2,11] Eric S. Lander,[1,12,13]† Job Dekker[5]†

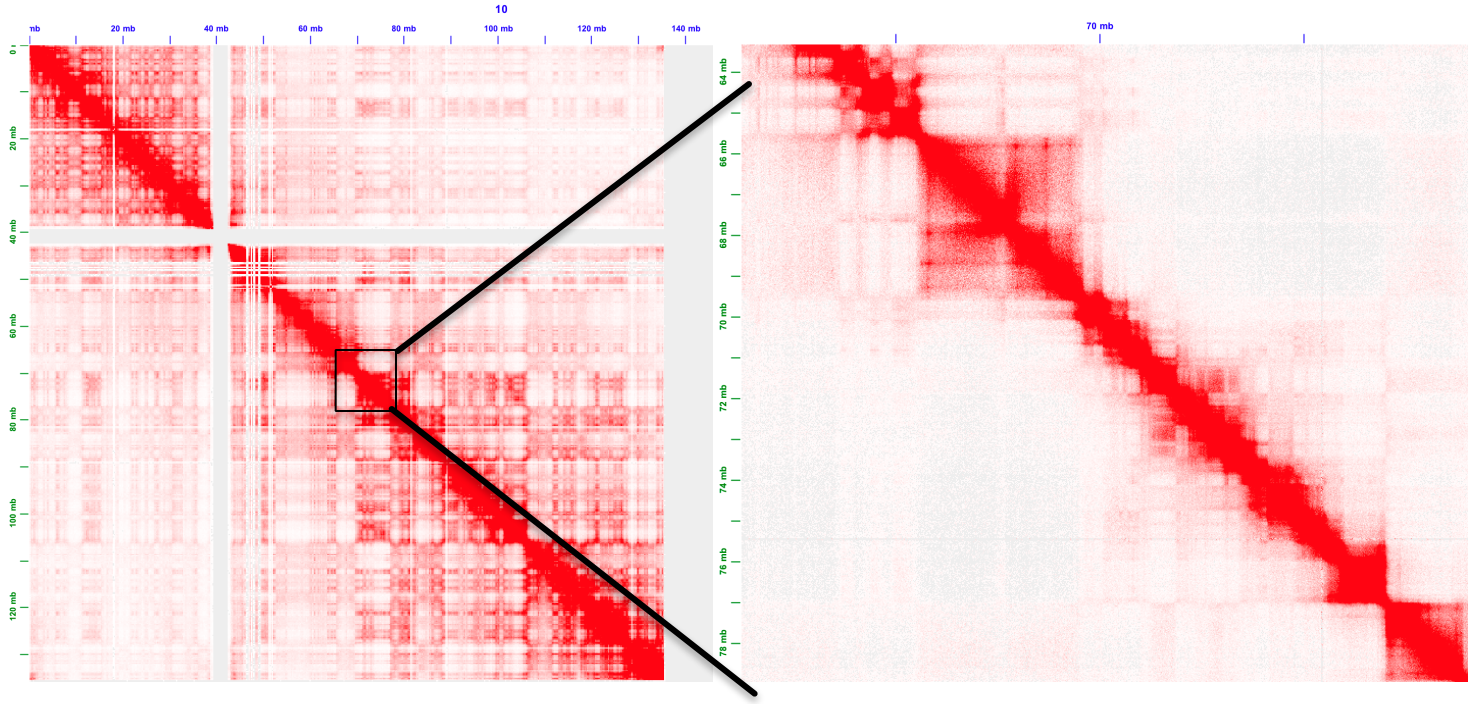
SCIENCE VOL 326 9 OCTOBER 2009



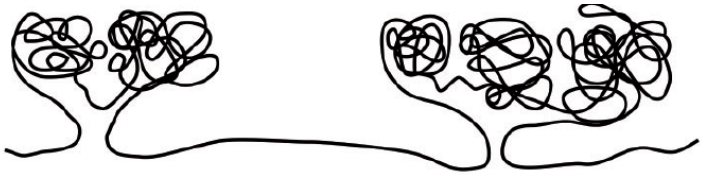


Data:
Rao et al. Aiden, Cell 2014

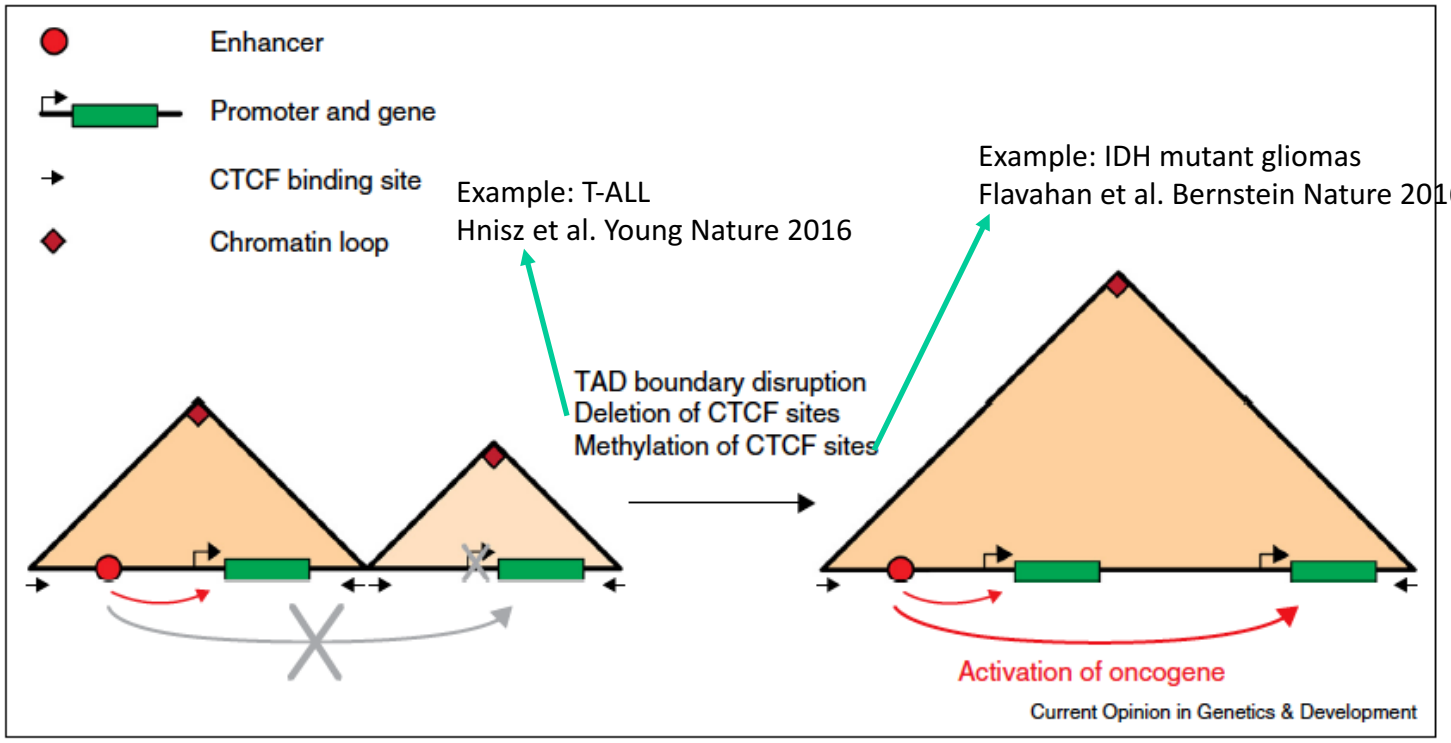
Topologically associating domains (TADs)



TADs have apparent hierarchical organization

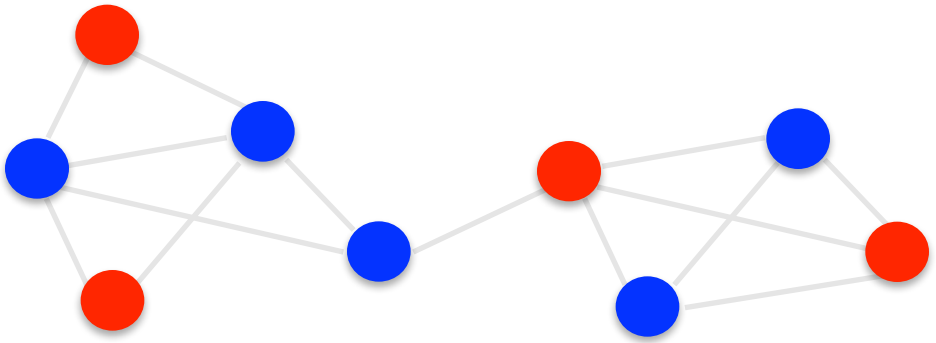


Local TAD boundary disruption activates oncogene



Valton and Dekker Curr. Opin. Genetics and Development 2016

Network modularity



$$Q \approx 0$$

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

adjacency matrix

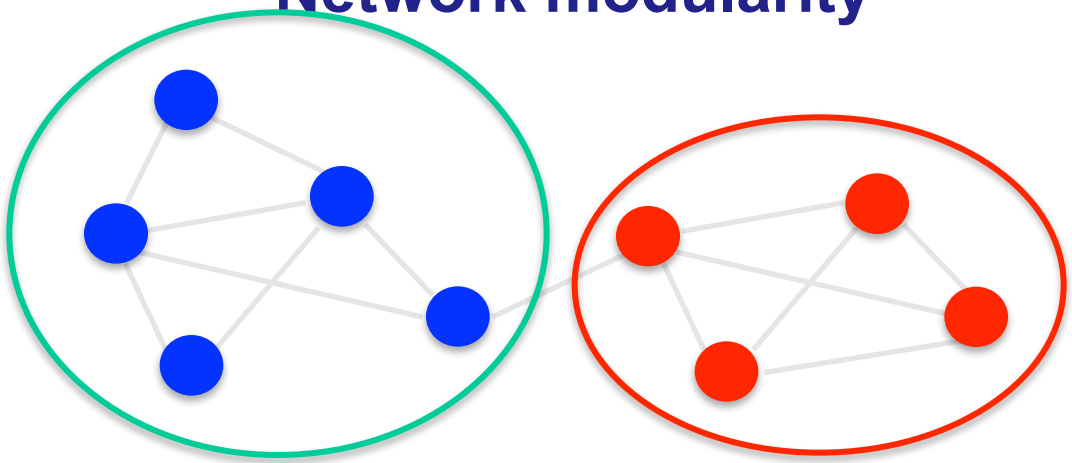
degree of i

whether or not i, j are in the same module

number of edges

expected number of edges between i and j

Network modularity



$$Q = Q_{max}$$

Optimization problem

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

adjacency matrix

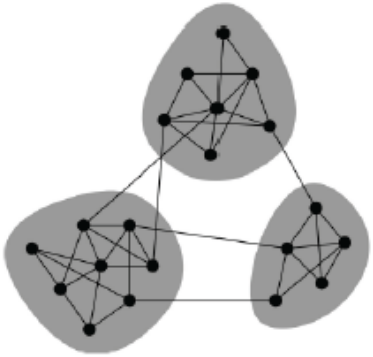
degree of i

whether or not i, j are in the same module

number of edges

expected number of edges between i and j

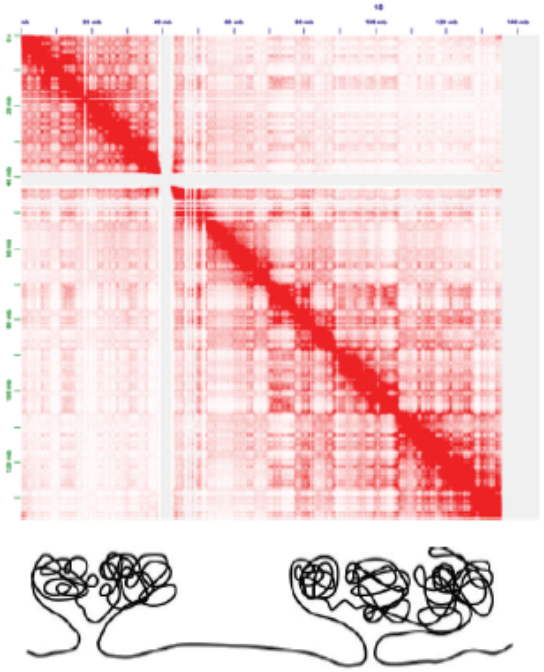
Identifying TADs in multiple resolutions



Modularity maximization

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

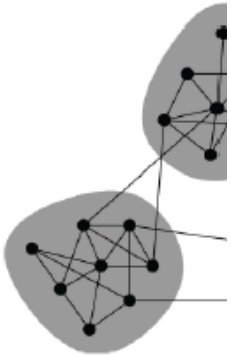
network	contact map
node	chromosome bin
edge	Hi-C contact
# of connections	coverage
module	domain



schematic adapted from ref. [2]

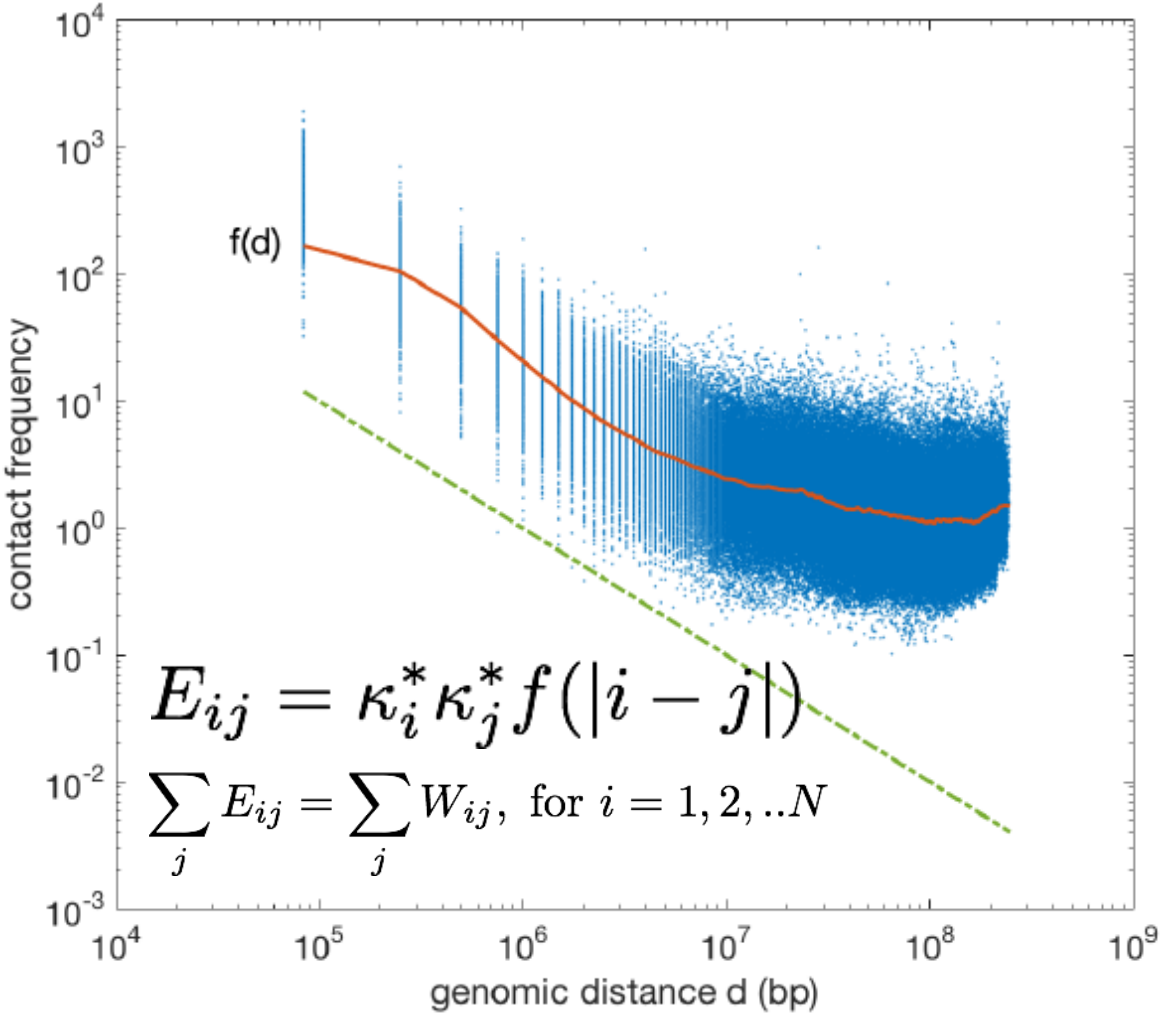
[Yan et al., *PLOS Comp. Bio.* (in revision, ‘17); bioRxiv 097345]

Identifying TADs in multiple resolutions



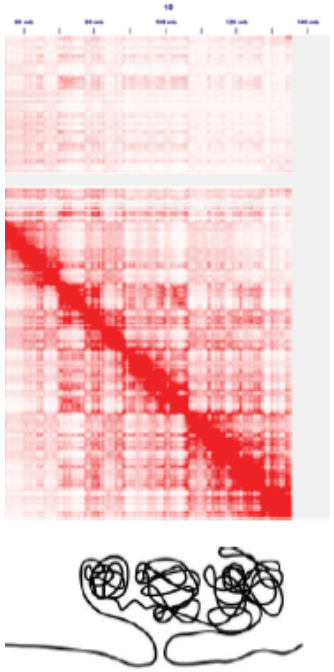
Modularity m

$$Q = \frac{1}{2m} \sum_{i,j}$$



$$E_{ij} = \kappa_i^* \kappa_j^* f(|i-j|)$$

$$\sum_j E_{ij} = \sum_j W_{ij}, \text{ for } i = 1, 2, ..N$$

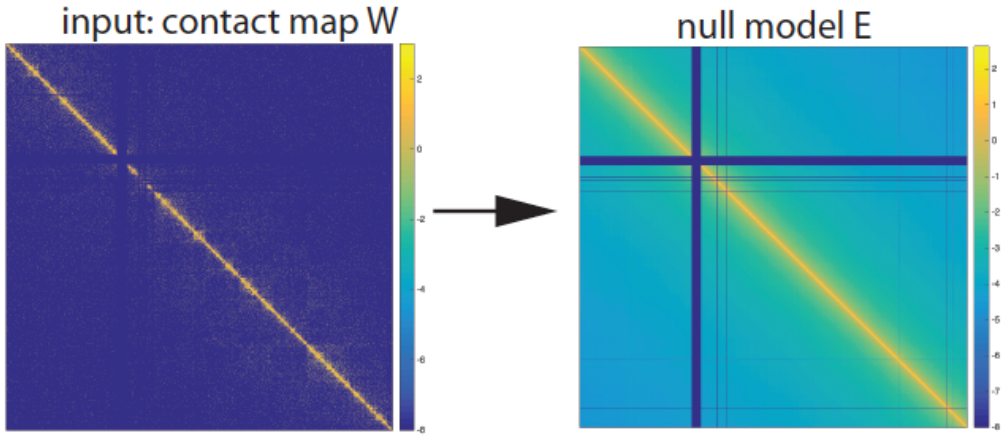


: adapted from ref. [2]

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

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

Identifying TADs in multiple resolutions



$$E_{ij} = \kappa_i^* \kappa_j^* f(|i-j|)$$

Numerically solve for κ_i^* in equations

$$\sum_j E_{ij} = \sum_j W_{ij}, \text{ for } i = 1, 2, ..N$$

Choose a particular resolution γ
Optimize Q over all possible partitions

$$Q = \frac{1}{2N} \sum_{ij} (W_{ij} - \gamma E_{ij}) \delta_{\sigma_i \sigma_j}$$ γ: resolution parameter

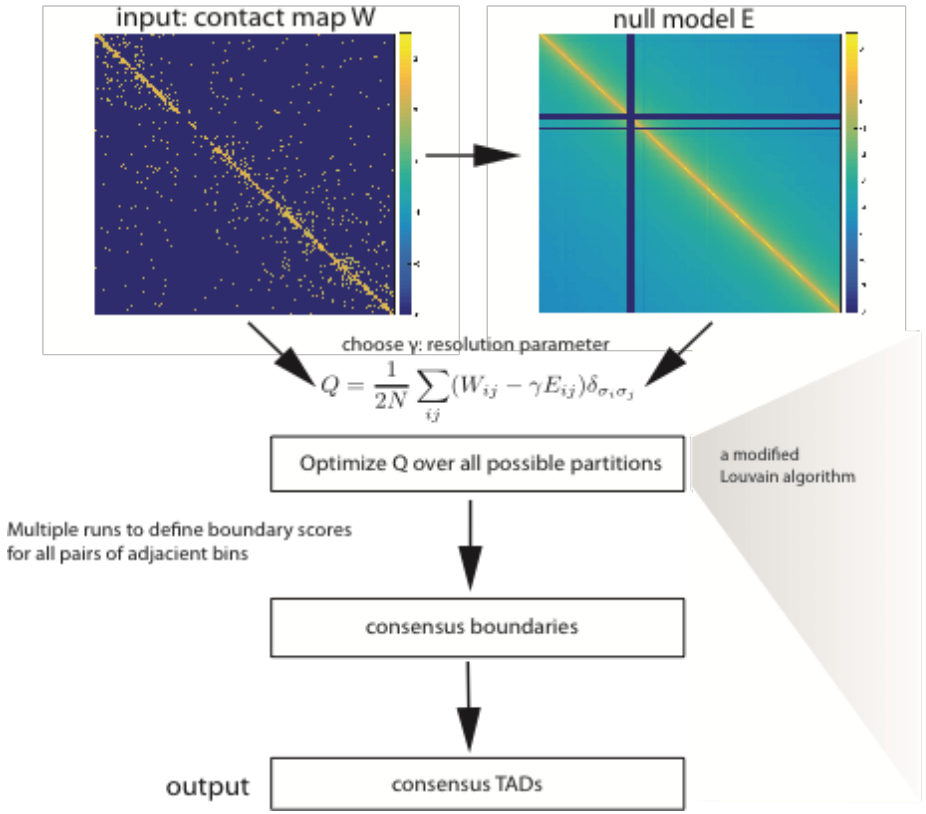
Multiple runs to define boundary scores for all pairs of adjacient bins

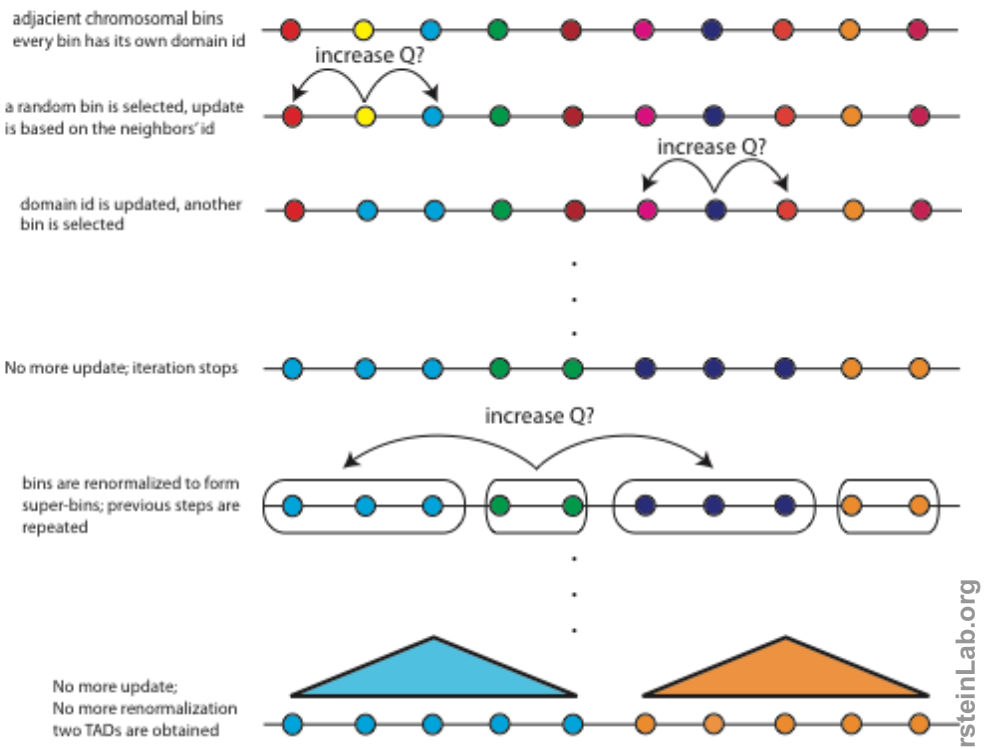
consensus boundaries based on the boundary scores

consensus TADs → output

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

Identifying TADs in multiple resolutions

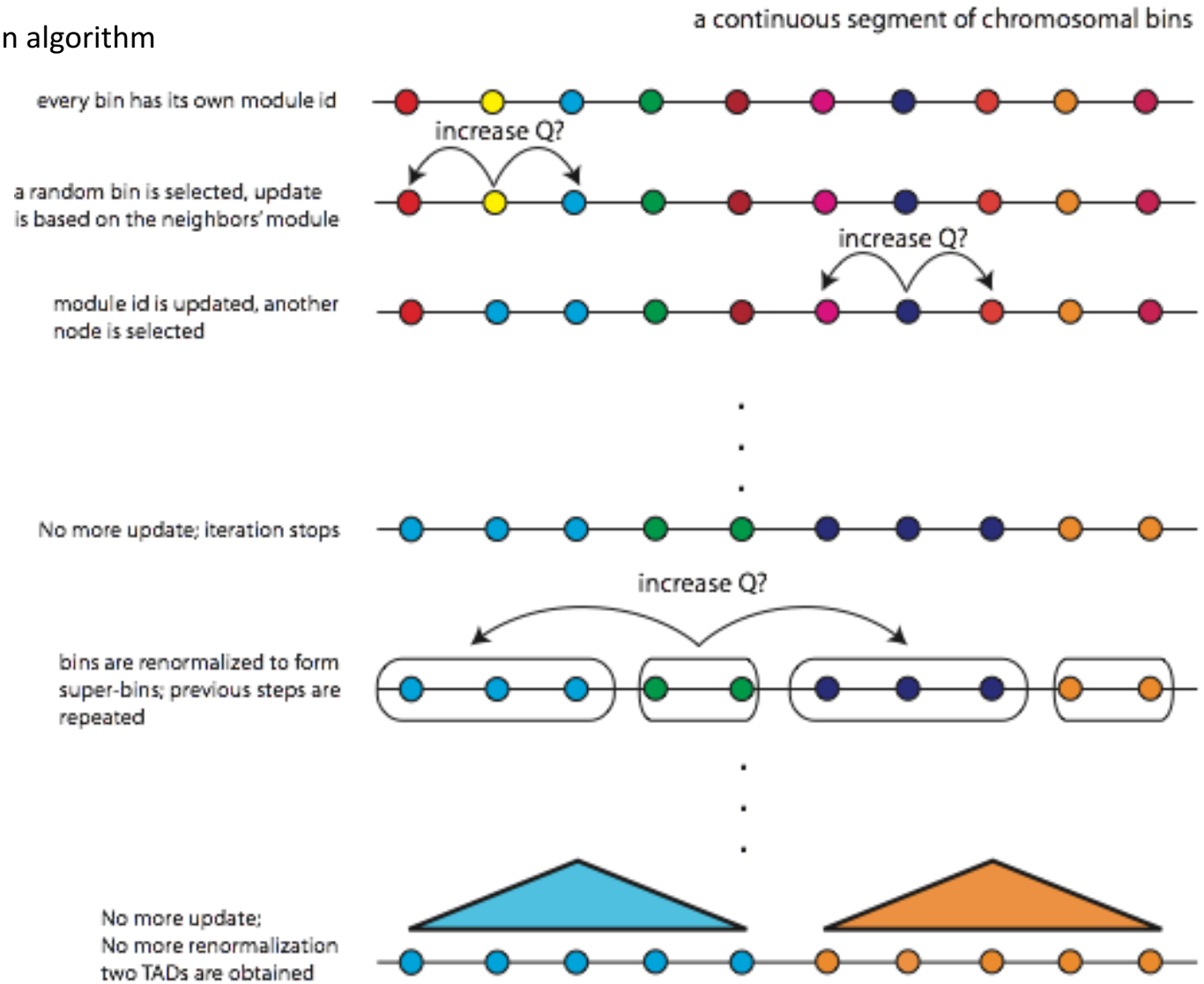




[Yan et al., *PLOS Comp. Bio.* (in revision, ‘17); bioRxiv 097345]

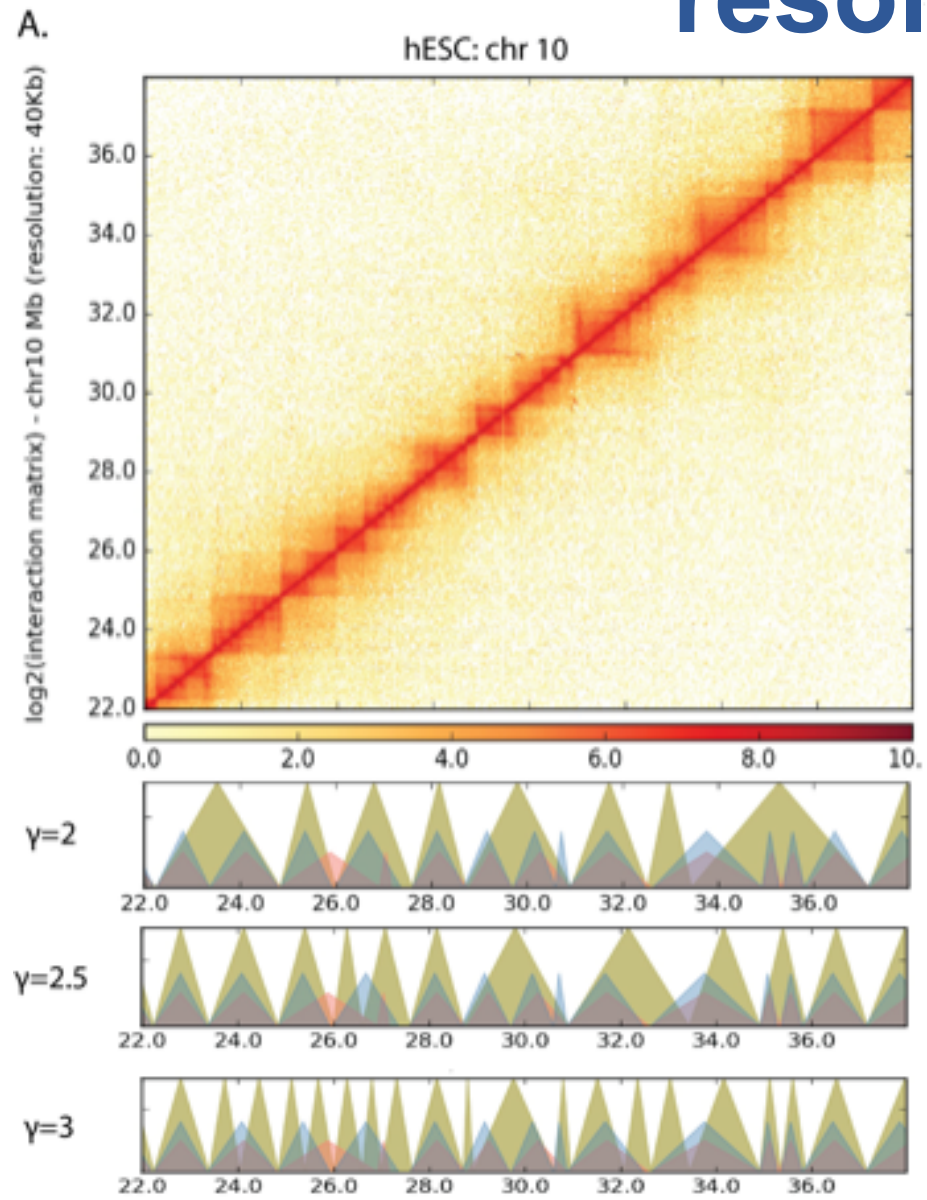
Identifying TADs in multiple resolutions

a modified Louvain algorithm

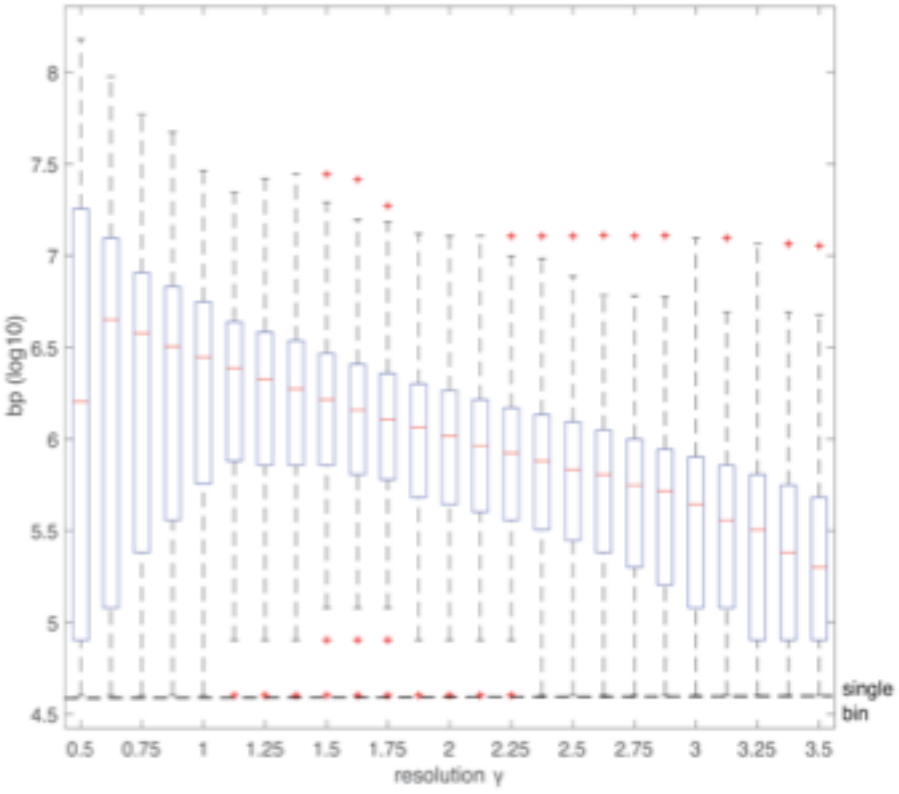


Identifying TADs in multiple resolutions

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



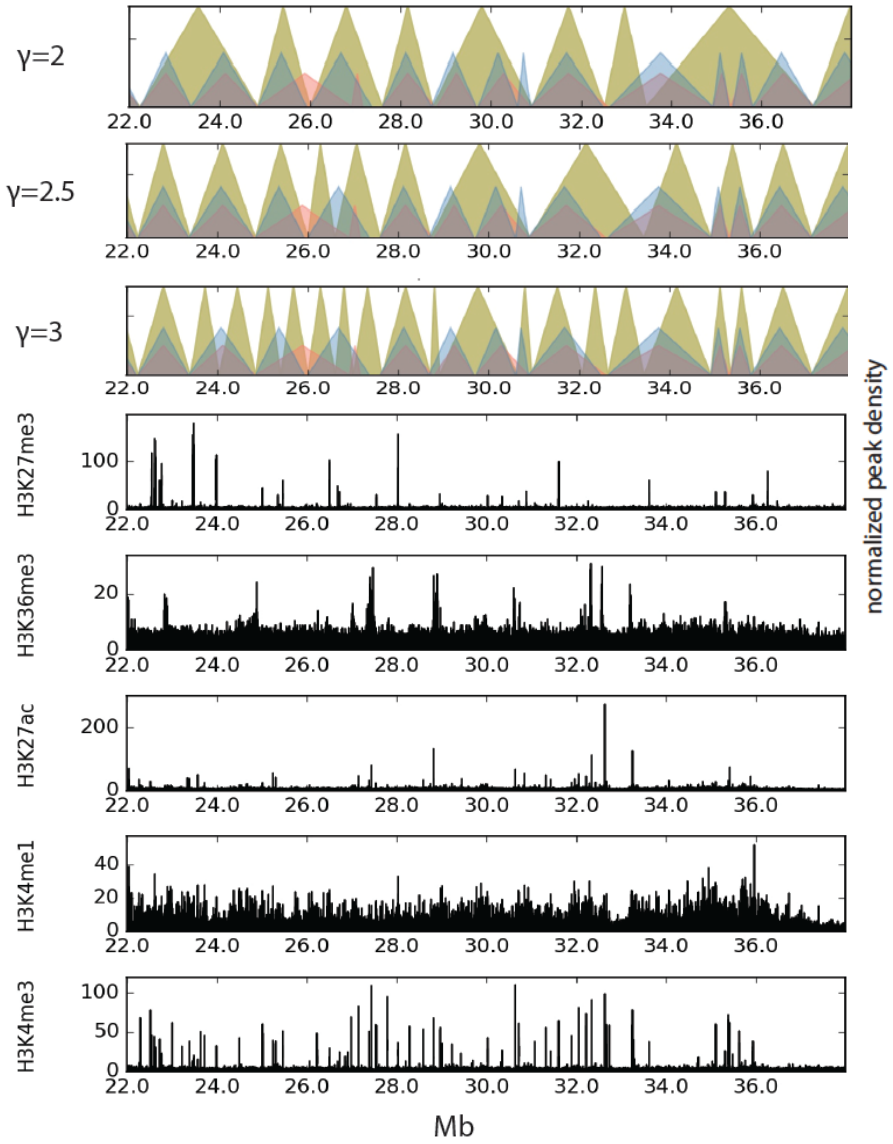
smaller TADs but are detected as the resolution increases

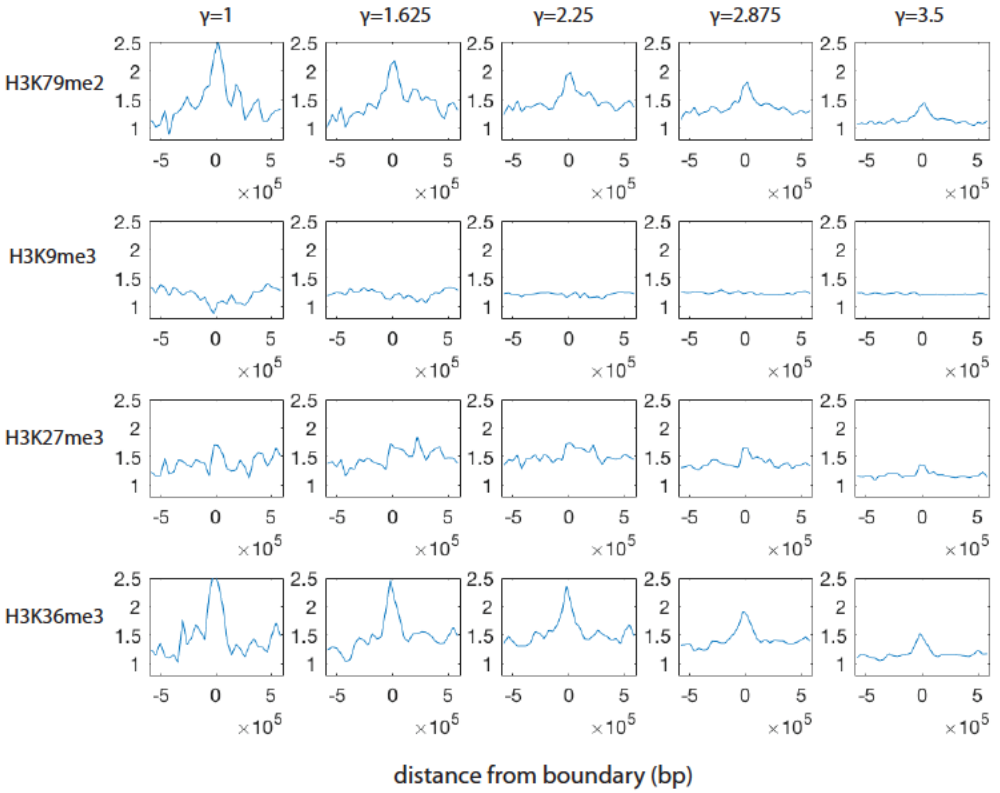


Evolution of Element Annotation, from Calling ChIP Peaks to Determining Genome Folding

- Characterizing Regulatory Sites on the Linear Genome
  - Original peak calling approach (with PeakSeq)
  - New Multi-scale "site" calling (with Music)
- Characterizing TADs from 3D Genome Folding
  - Using modularity for identification, at multiple scales (with MrTADFinder)
  - Developing an appropriate null expectation
- Features of Multi-resolution TADs
  - Specific TFs & HMs associated with TAD boundaries at different scales
  - Assoc. strong enough to build a predictor
  - HOT regions at boundaries
  - Relation to somatic mutations
- Technical Analysis of TADs
  - Spectral analysis quantifying reproducibility of Hi-C data sets (with HiC-Spector)

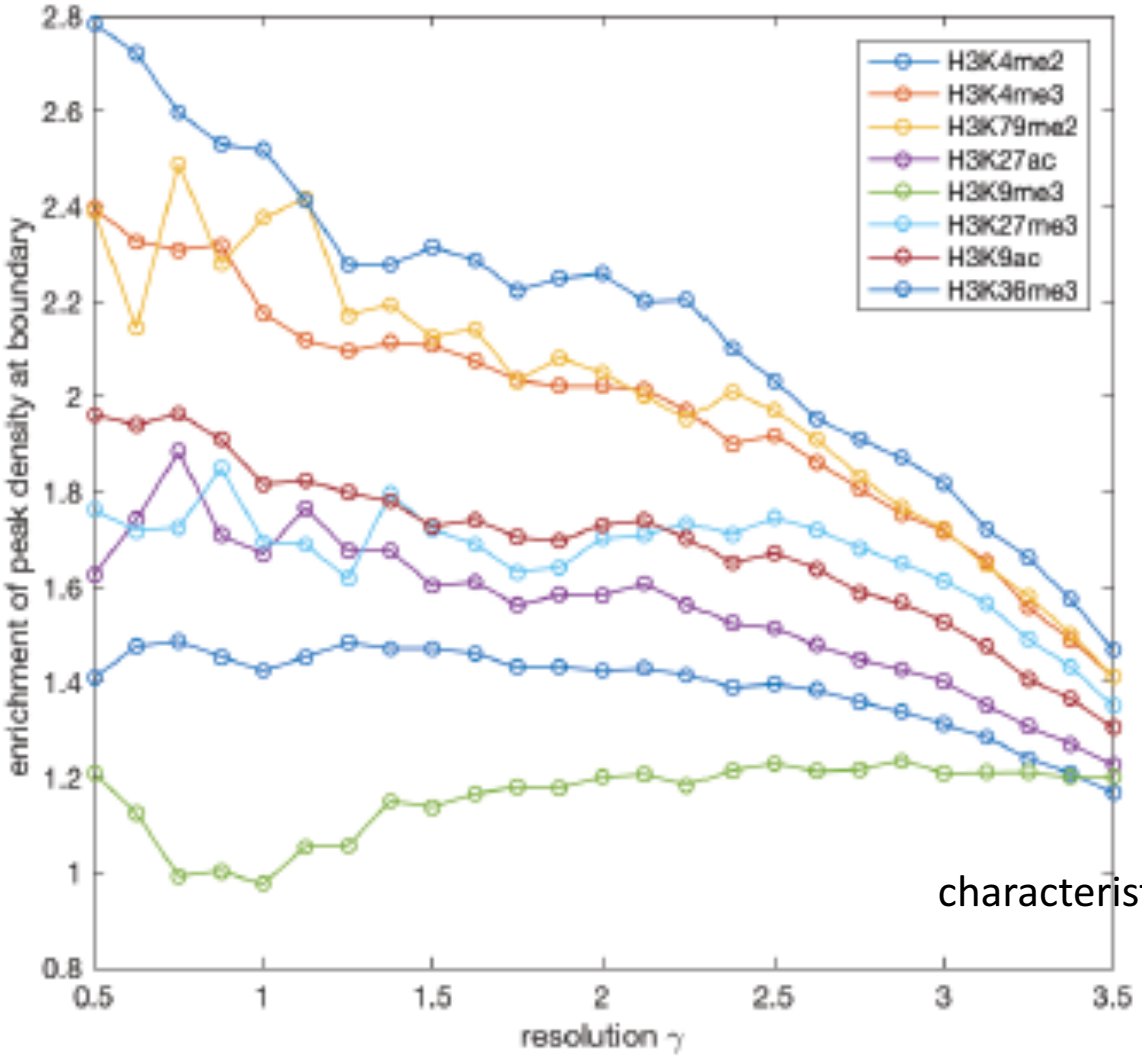
Enrichment of histone features at different resolution





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

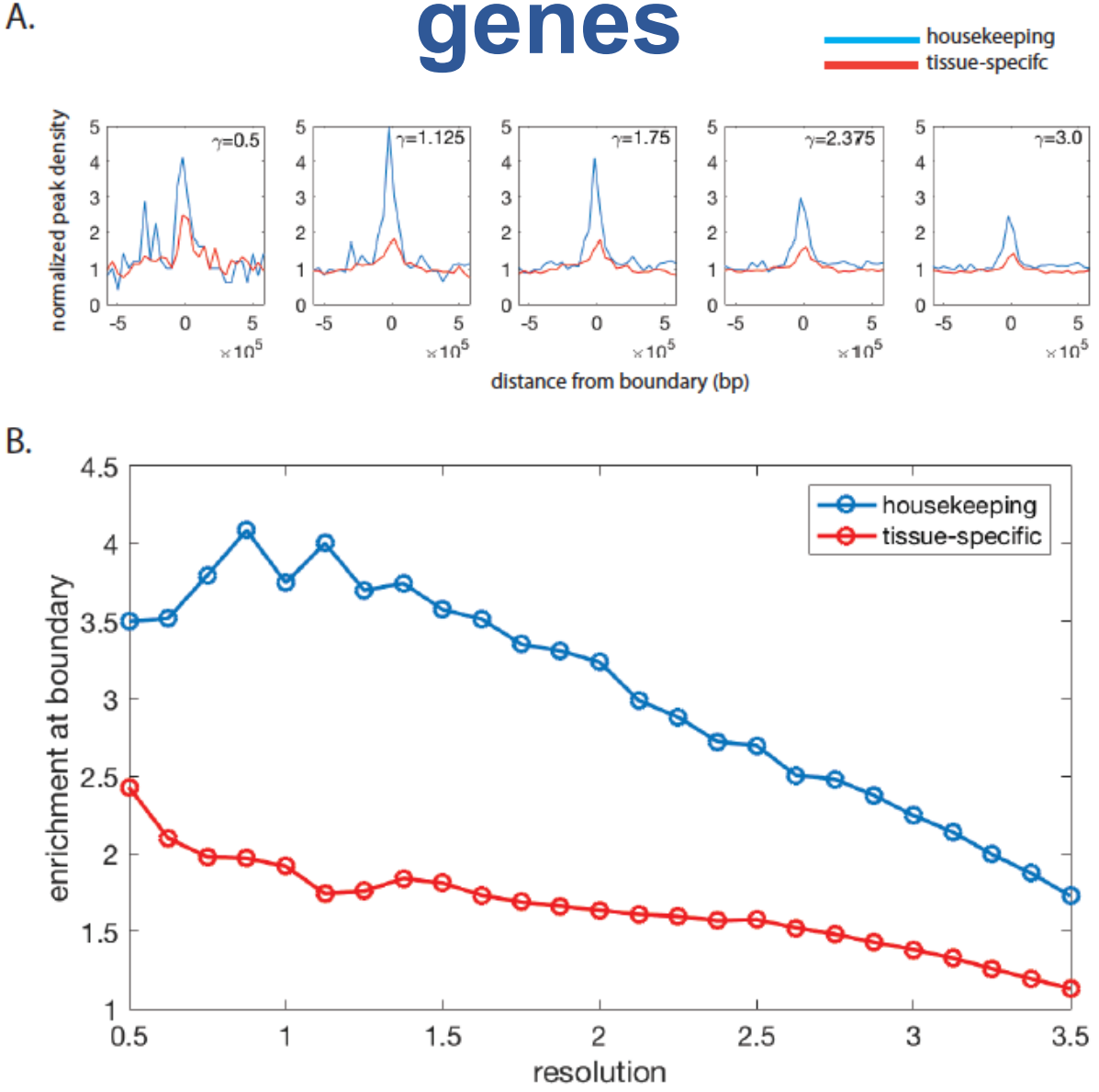
Enrichment of histone features at different resolution



characteristic length scale

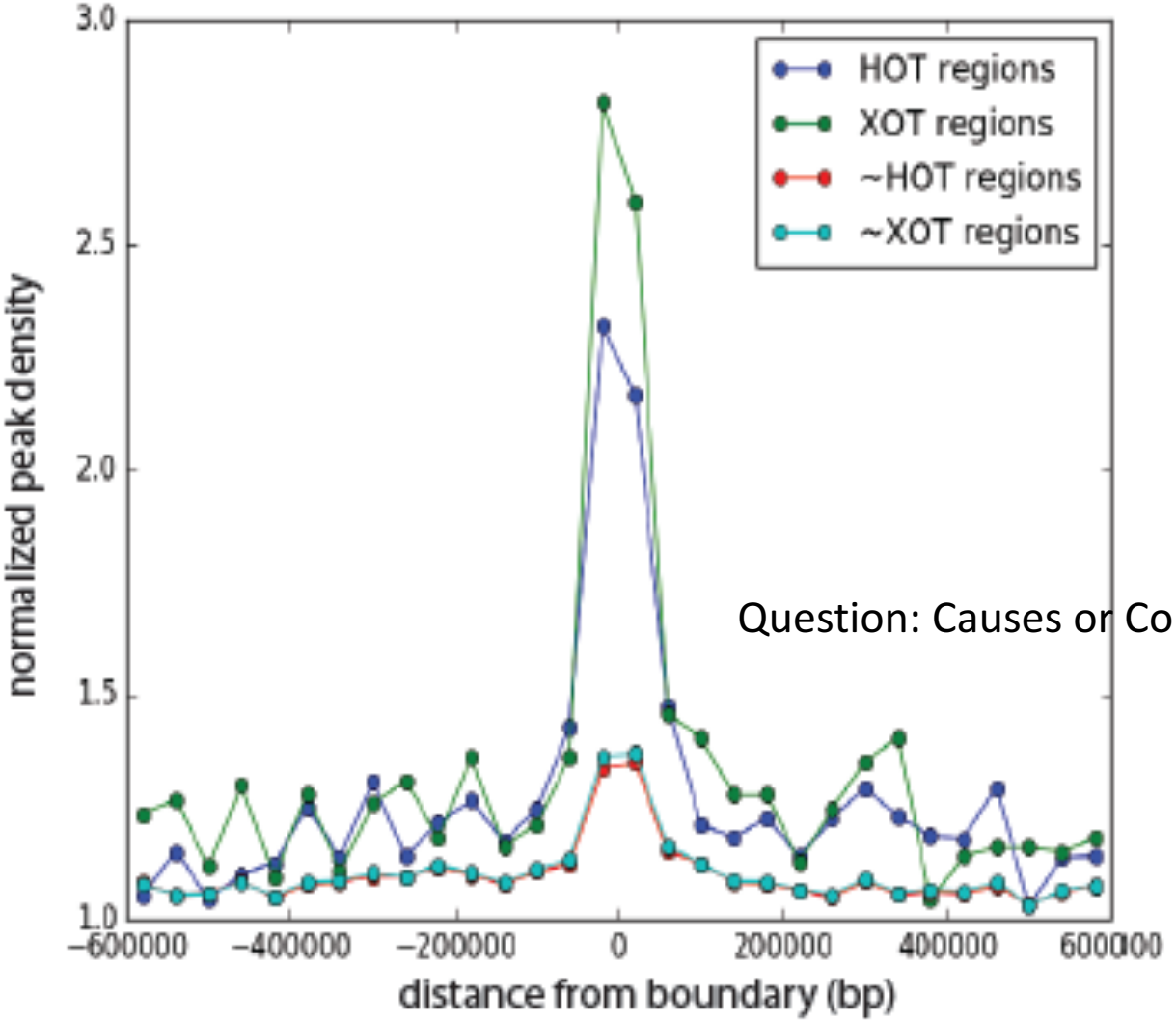
[Yan et al., *PLOS Comp. Bio.* (in revision, '17); bioRxiv 097345]

House-keeping vs tissue-specific genes



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

Enrichment of TF binding sites near boundaries

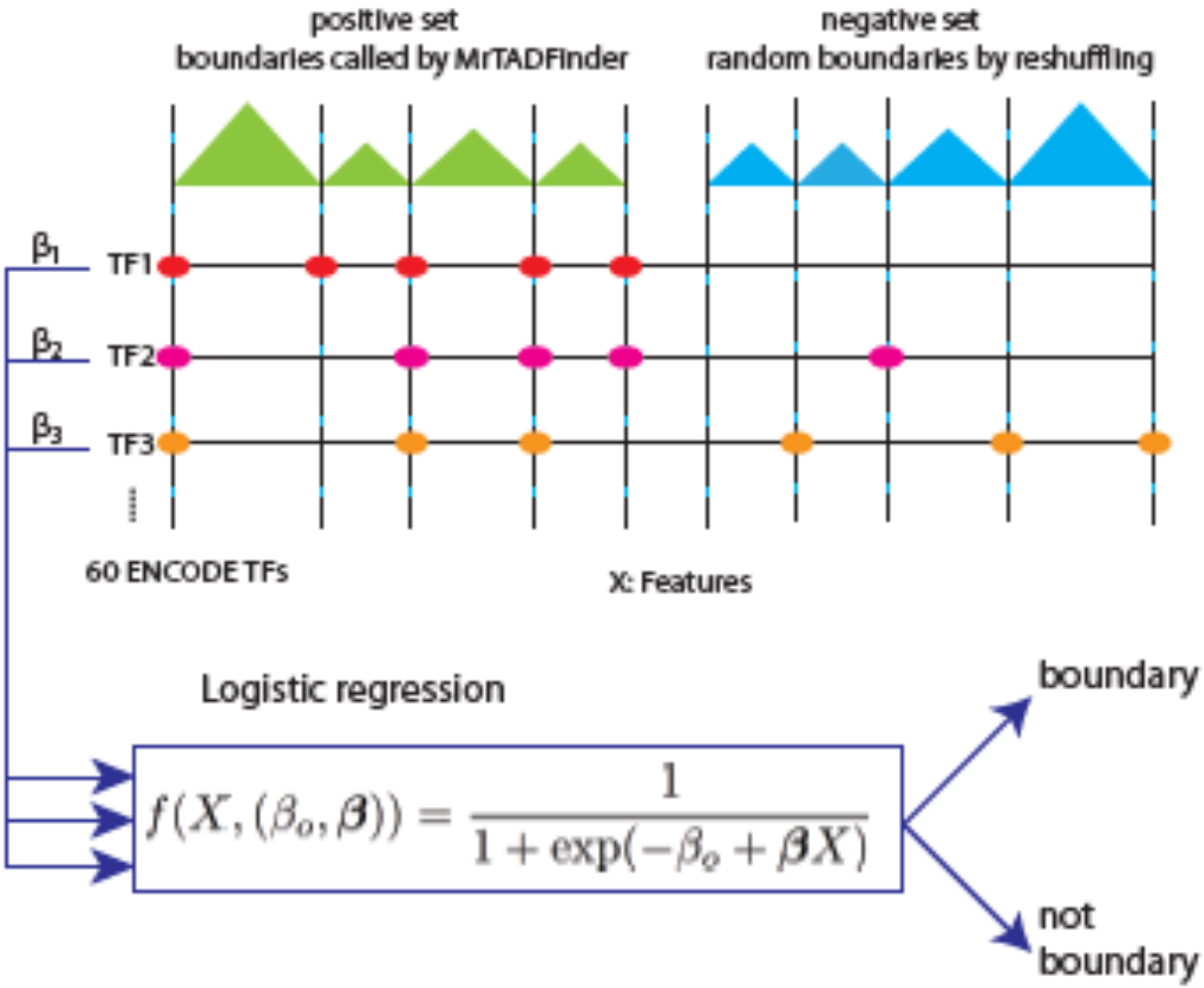


Question: Causes or Consequences?

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

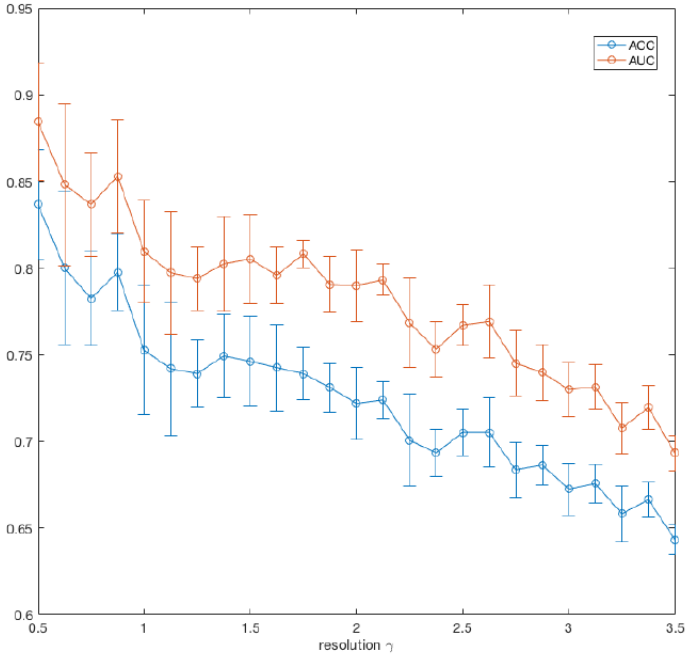
Predicting TAD boundaries using TFs binding pattern

Classification problem:



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

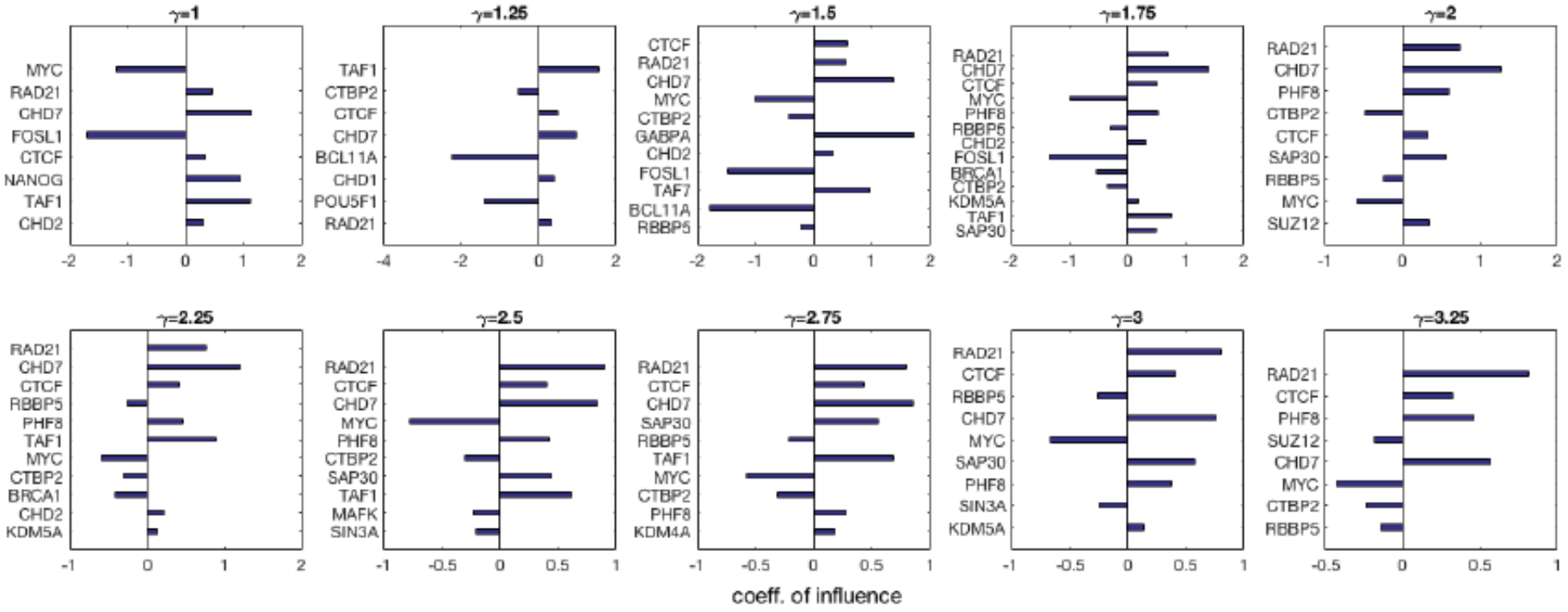
model performance



Predicting TAD boundaries using chromatin features

Which transcription factors play a role in border formation?

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



contribution of individual factors

Domain organization shapes mutational landscape

LETTER

OPEN

doi:10.1038/nature14221

Cell-of-origin chromatin organization shapes the mutational landscape of cancer

Paz Polak[1,2]*, Rosa Karlić[3]*, Amnon Koren[2,4], Robert Thurman[5], Richard Sandstrom[5], Michael S. Lawrence[2], Alex Reynolds[5], Eric Rynes[5], Kristian Vlahoviček[3,6], John A. Stamatoyannopoulos[5,7] & Shamil R. Sunyaev[1,2]

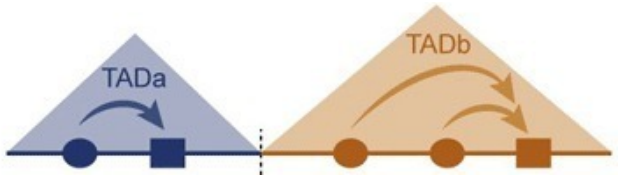
Cancer is a disease potentiated by mutations in somatic cells. Cancer mutations are not distributed uniformly along the human genome. Instead, different human genomic regions vary by up to fivefold in the local density of cancer somatic mutations[1], posing a fundamental problem for statistical methods used in cancer genomics. Epigenomic organization has been proposed as a major determinant of the cancer mutational landscape[1–5]. However, both somatic mutagenesis and epigenomic features are highly cell-type-specific[6,7]. We investigated the distribution of mutations in multiple independent samples of diverse cancer types and compared them to cell-type-specific epigenomic features. Here we show that chromatin accessibility and modification, together with replication timing, explain up to 86% of the variance in mutation rates along cancer genomes. The best predictors of local somatic mutation density are epigenomic features derived from the most likely cell type of origin of the corresponding malignancy. Moreover, we find that cell-of-origin chromatin features are

EPIGENOME ROADMAP
A *Nature* special issue
nature.com/epigenomeroadmap

cell types from 45 different tissue types, encompassing the established or likely cell types of origin of most of the cancer types that we investigated (Methods and Extended Data Fig. 2). Notably, these data derive from primary human cells and tissues rather than malignant cell lines. These epigenetic features comprised eight different types of variables, including DNase I hypersensitivity (a global measure of chromatin accessibility)[7] and various histone modifications. An example of the variation in mutation density along chromosomes at a 1 Mb scale together with the density of DNase I hypersensitive sites (DHSs) is shown in Fig. 1. In this case, as in most other cases (see later), epigenomics features indicative of active chromatin and transcription were associated with low mutation density, whereas repressive chromatin features were associated with regions of high mutation density. Notably, these stat-

Domain organization shapes mutational landscape

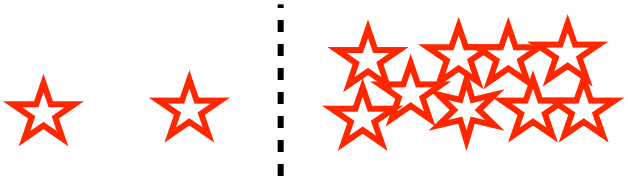
TADs identified in MCF7



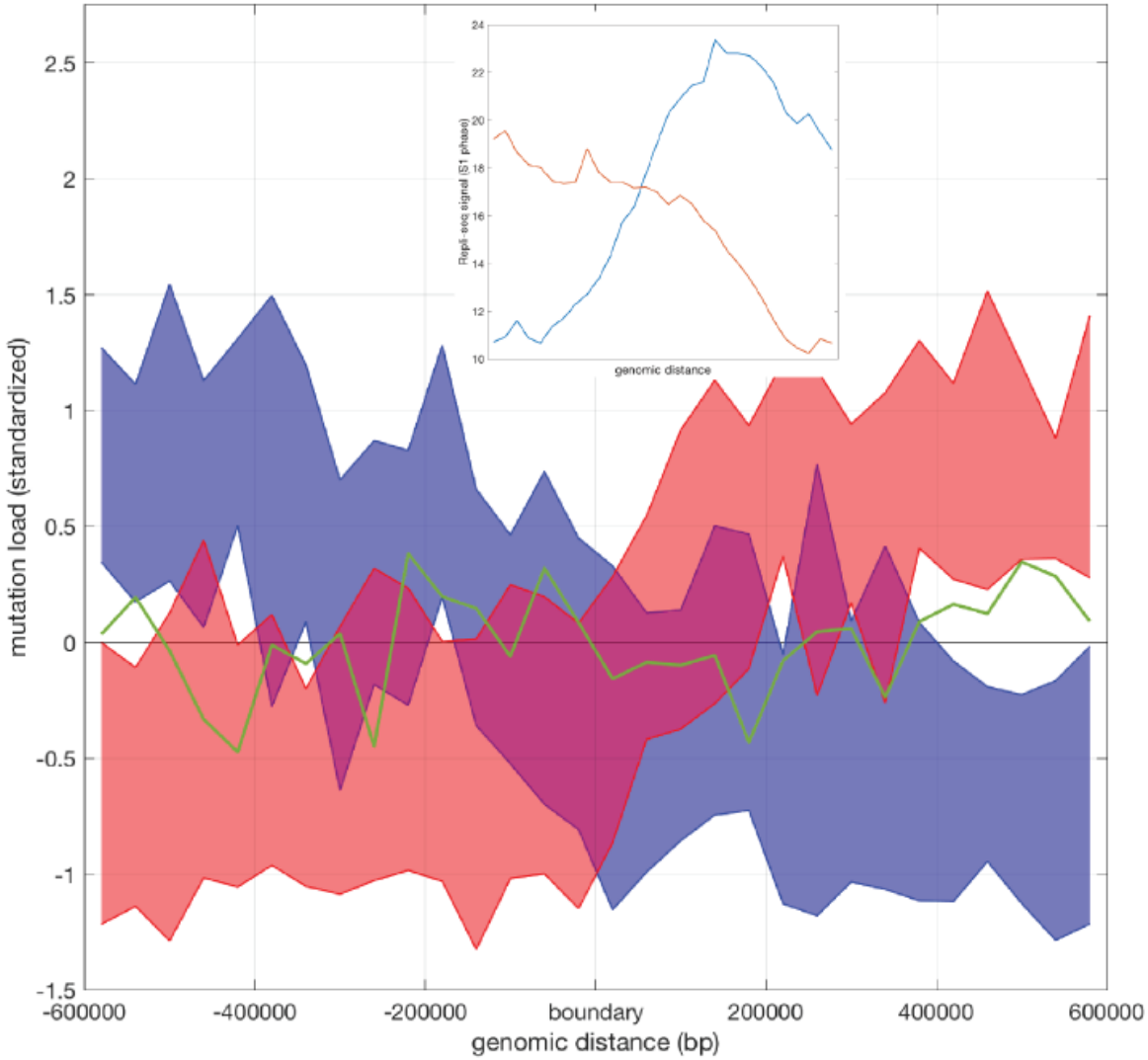
mutations called from breast cancer samples (~700 donors)

Cluster 1

Cluster 2

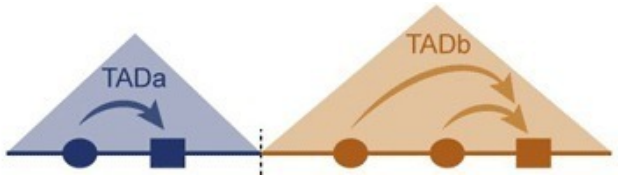


[Yan et al., *PLOS Comp. Bio.* ('17)]



Domain organization shapes mutational landscape

TADs identified in MCF7



mutations called from breast cancer samples

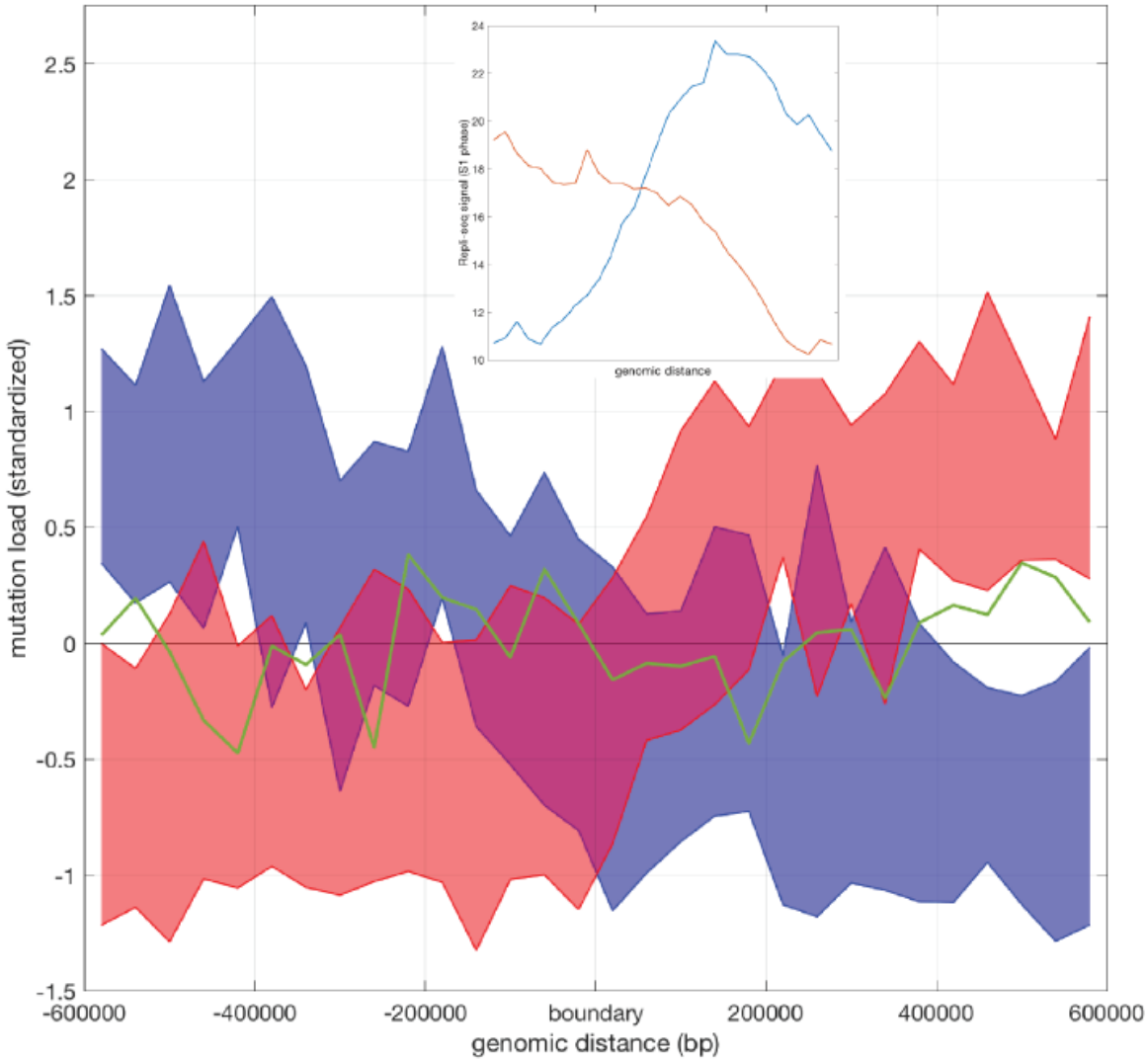
Cluster 1

late | early

Cluster 2

early | late

[Yan et al., *PLOS Comp. Bio.* ('17)]



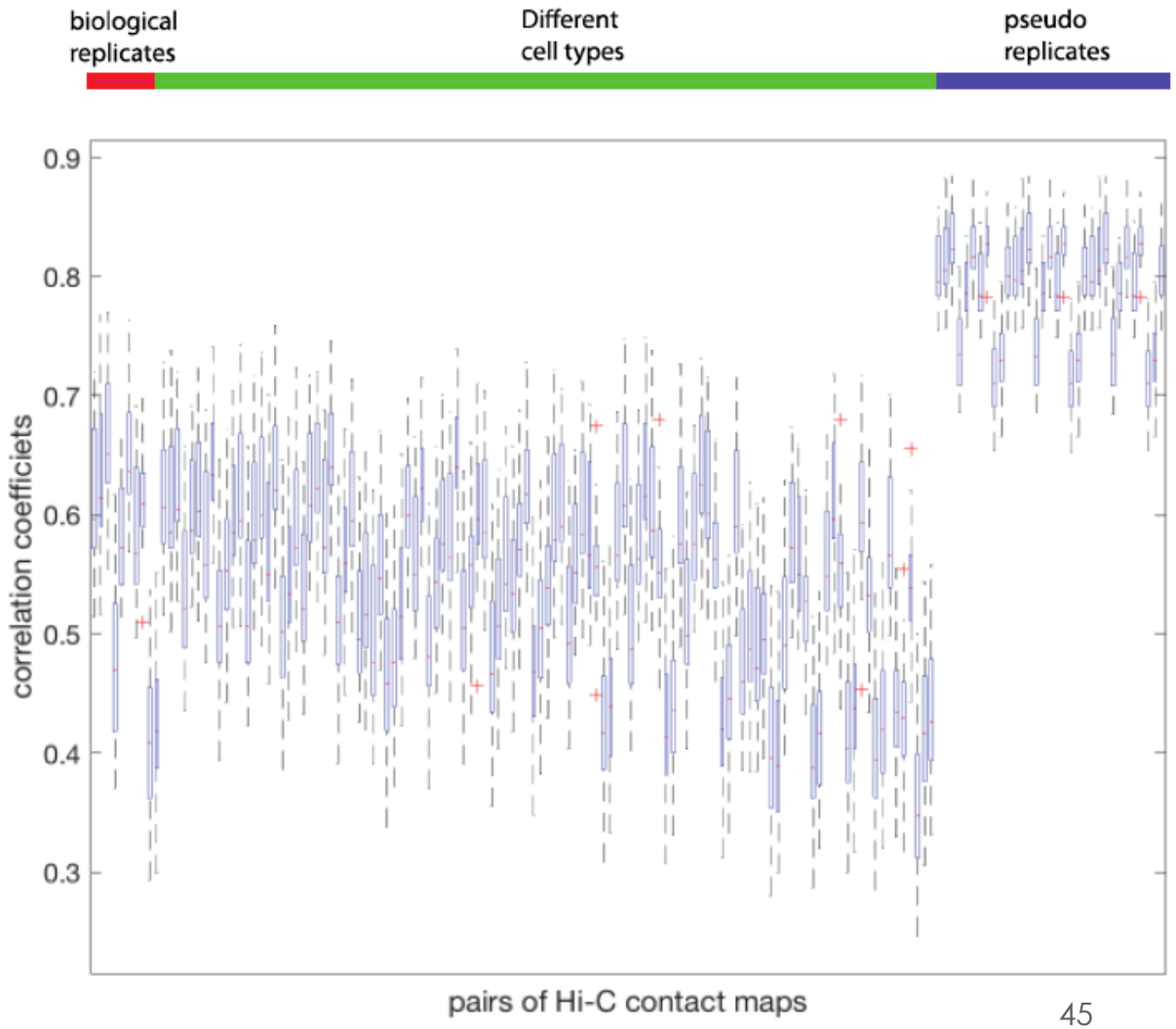
Evolution of Element Annotation, from Calling ChIP Peaks to Determining Genome Folding

- Characterizing Regulatory Sites on the Linear Genome
  - Original peak calling approach (with PeakSeq)
  - New Multi-scale "site" calling (with Music)
- Characterizing TADs from 3D Genome Folding
  - Using modularity for identification, at multiple scales (with MrTADFinder)
  - Developing an appropriate null expectation
- Features of Multi-resolution TADs
  - Specific TFs & HMs associated with TAD boundaries at different scales
  - Assoc. strong enough to build a predictor
  - HOT regions at boundaries
  - Relation to somatic mutations
- Technical Analysis of TADs
  - Spectral analysis quantifying reproducibility of Hi-C data sets (with HiC-Spector)

Quantifying reproducibility of Hi-C data

ENCODE Hi-C data

Tissue/Morphology	cell type	# interactions (millions)
Lung/Epithelial	A549	33
		30
Kidney/Epithelial	Caki2	36
		47
Kidney/Epithelial	G401	61
		53
Prostate/Epithelial	LNCaP	18
		15
Lung/Epithelial	NCI-H460	42
		29
Pancreas/Epithelial	Panc1	37
		51
Skin/Epithelial	RPMI-7951	32
		49
Skin/Stellate	SK-MEL-5	46
		11
Brain/Epithelial	SK-N-DZ	16
		10
Brain/Epithelial	SK-N-MC	25
		13
Mammary Gland/Epitheliа	T47D	34
		36



[Yan et al., *Bioinformatics* ('17)]

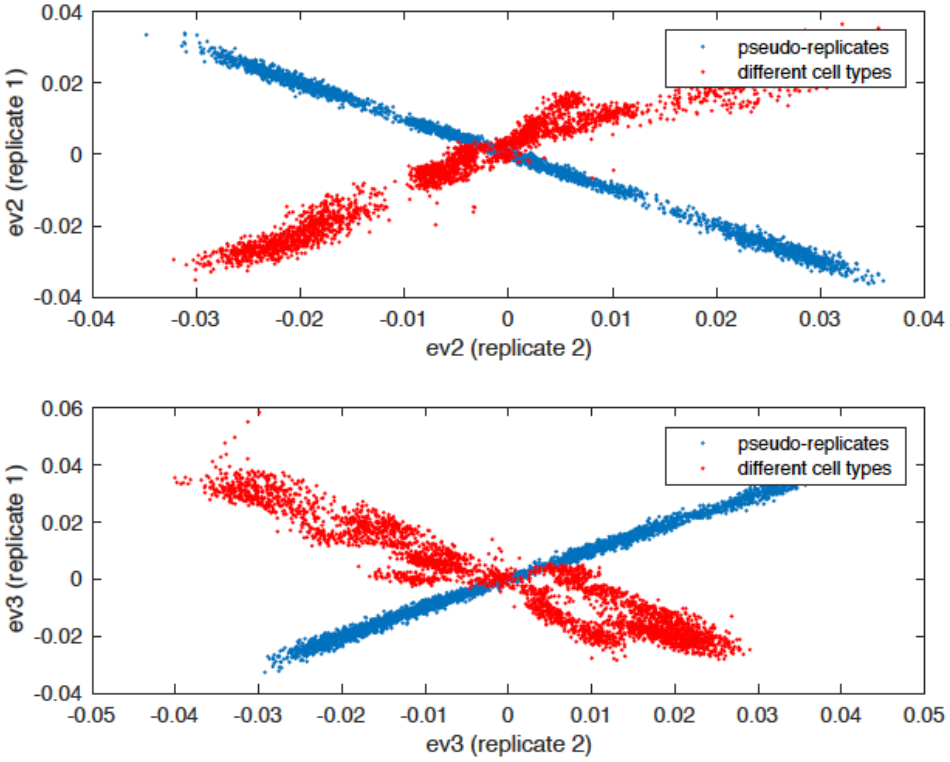
Quantifying reproducibility of Hi-C data

Is there a better way to decompose the contact map W (matrix)?

- Spectral clustering commonly used in image processing
- Transform W into the Laplacian matrix

$$\mathcal{L} = I - D^{-1/2} W D^{-1/2}, D_{ii} = \sum_j W_{ij}$$

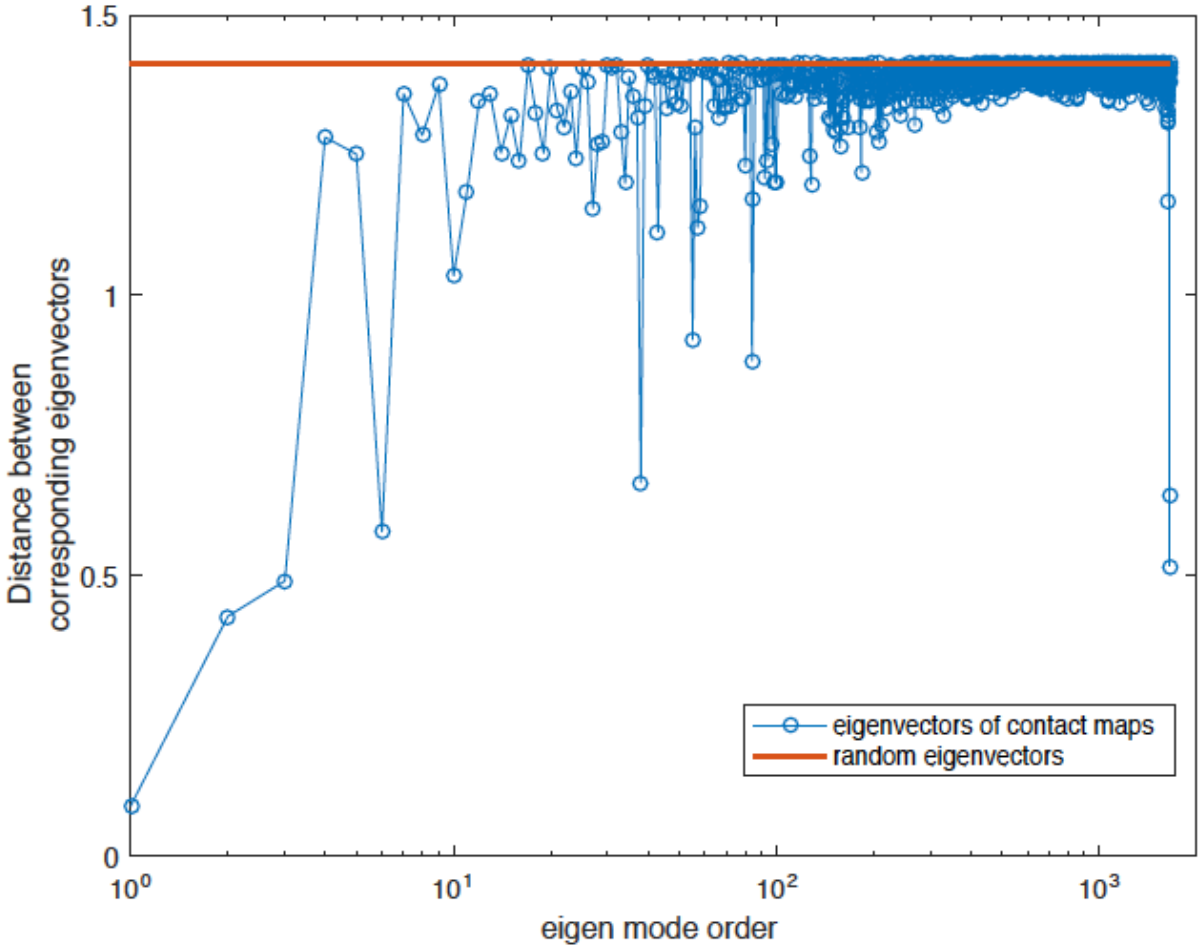
- Decomposed into eigenvectors, and consider only the leading ones (dimension reduction)
- Distance between the corresponding vectors



Yan KK et al. Bioinformatics 2017

Quantifying reproducibility of Hi-C data

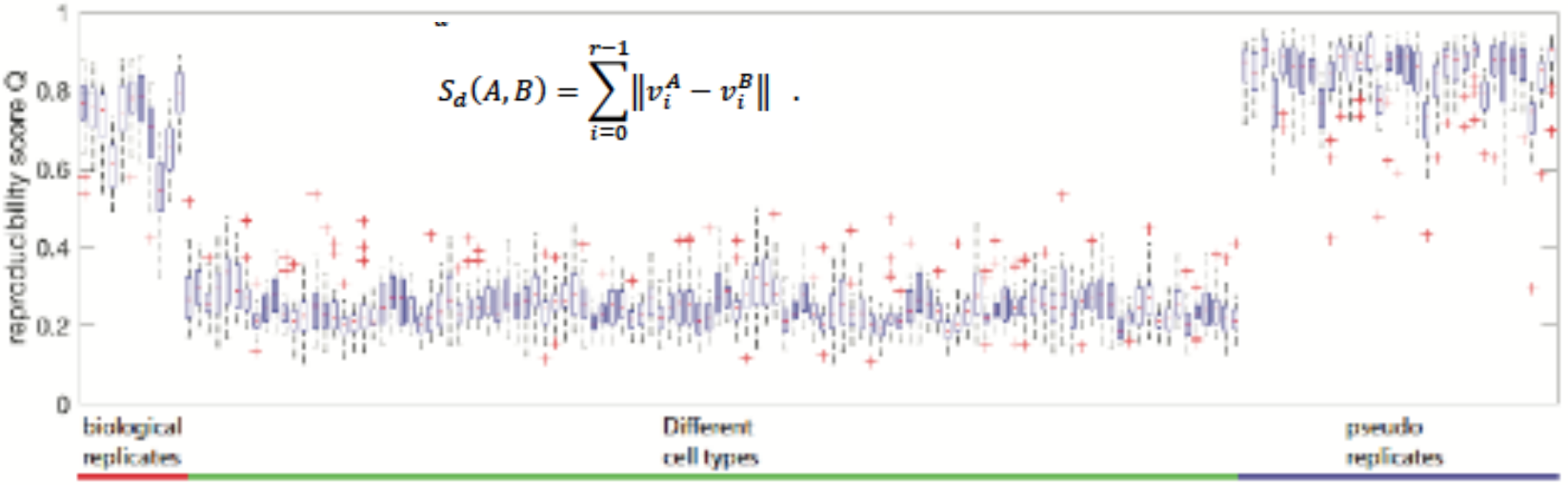
How many eigenvectors should be used?

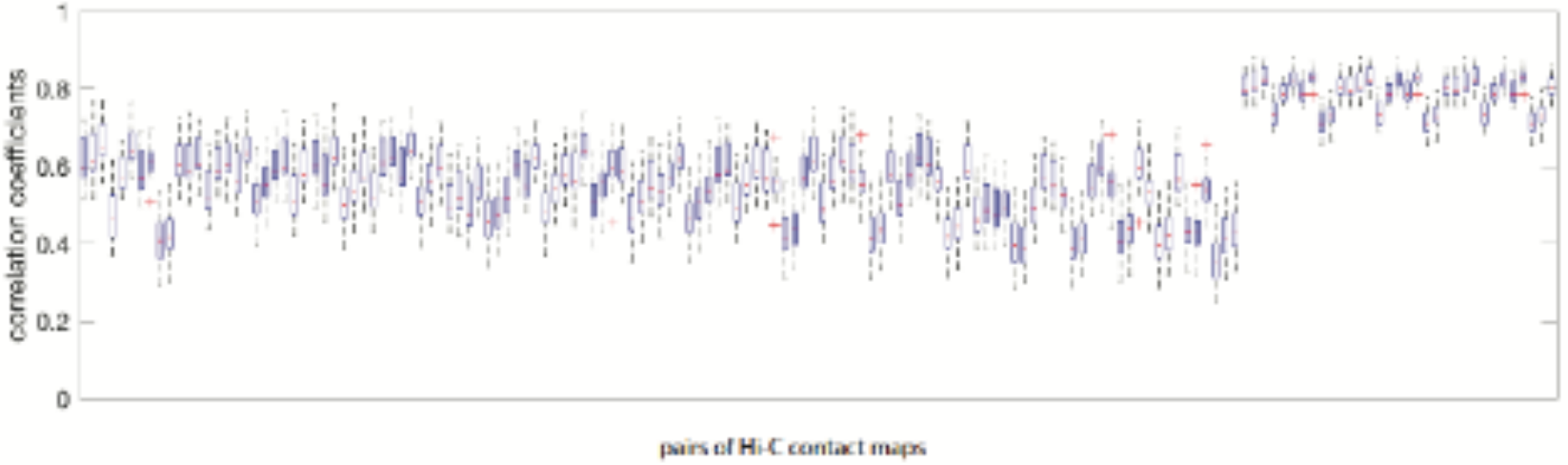


Yan KK et al. Bioinformatics 2017

Quantifying reproducibility of Hi-C data

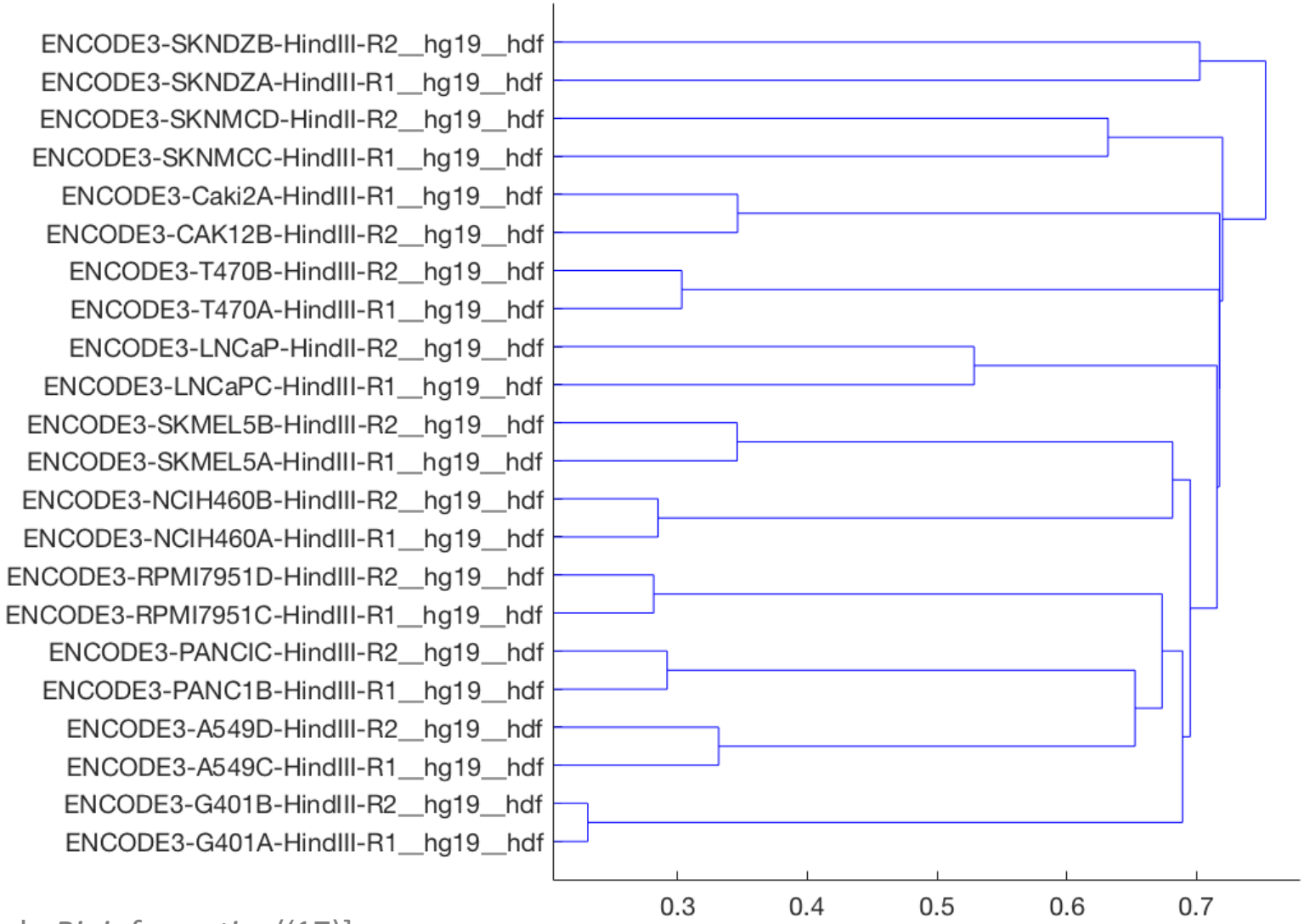
$$S_d(A,B) = \sum_{i=0}^{r-1} \|v_i^A - v_i^B\| \quad .$$





Yan KK et al. Bioinformatics 2017

A distance measure between two contact maps



[Yan et al., *Bioinformatics* ('17)]

Evolution of Element Annotation, from Calling ChIP Peaks to Determining Genome Folding

- Characterizing Regulatory Sites on the Linear Genome
  - Original peak calling approach (with PeakSeq)
  - New Multi-scale "site" calling (with Music)
- Characterizing TADs from 3D Genome Folding
  - Using modularity for identification, at multiple scales (with MrTADFinder)
  - Developing an appropriate null expectation
- Features of Multi-resolution TADs
  - Specific TFs & HMs associated with TAD boundaries at different scales
  - Assoc. strong enough to build a predictor
  - HOT regions at boundaries
  - Relation to somatic mutations
- Technical Analysis of TADs
  - Spectral analysis quantifying reproducibility of Hi-C data sets (with HiC-Spector)

Evolution of Element Annotation,
from Calling ChIP Peaks to Determining Genome Folding

- Characterizing Regulatory Sites on the Linear Genome
  - Original peak calling approach (with PeakSeq)
  - New Multi-scale "site" calling (with Music)
- Characterizing TADs from 3D Genome Folding
  - Using modularity for identification, at multiple scales (with MrTADFinder)
  - Developing an appropriate null expectation
- Features of Multi-resolution TADs
  - Specific TFs & HMs associated with TAD boundaries at different scales
  - Assoc. strong enough to build a predictor
  - HOT regions at boundaries
  - Relation to somatic mutations
- Technical Analysis of TADs
  - Spectral analysis quantifying reproducibility of Hi-C data sets (with HiC-Spector)

MUSIC.gersteinlab.org - A **Harmanci**, J Rozowsky

github.com/gersteinlab/**MrTADfinder** - K **Yan**, S Lou

github.com/gersteinlab/**HiC-spector**

K **Yan**, G Gurkan Yardimci, C Yan, WS Noble

Hiring Postdocs. See **Jobs**.gersteinlab.org

Acknowledgments

Extra

Info about content in this slide pack

- General PERMISSIONS
  - This Presentation is copyright Mark Gerstein, Yale University, 2016.
  - Please read statement at www.**gersteinlab.org/misc/permissions.html** .
  - Feel free to use slides & images in the talk with PROPER acknowledgement (via citation to relevant papers or link to gersteinlab.org). Paper references in the talk are mostly from Papers.GersteinLab.org.

- PHOTOS & IMAGES. For thoughts on the source and permissions of many of the photos and clipped images in this presentation see streams.gerstein.info .