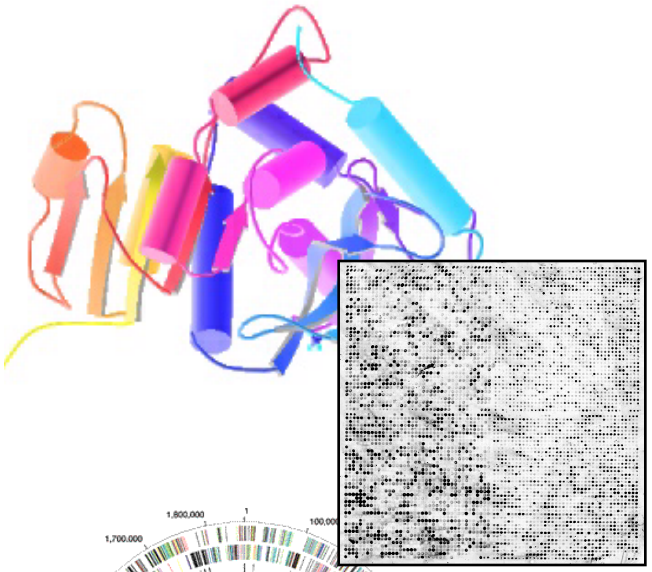
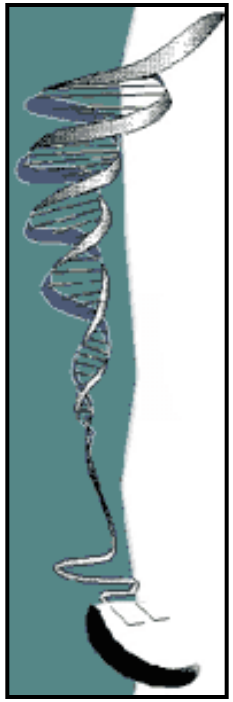
Biomed. Data Science:

Basic Multi-omic Analyses





Mark Gerstein, Yale University

gersteinlab.org/courses/452

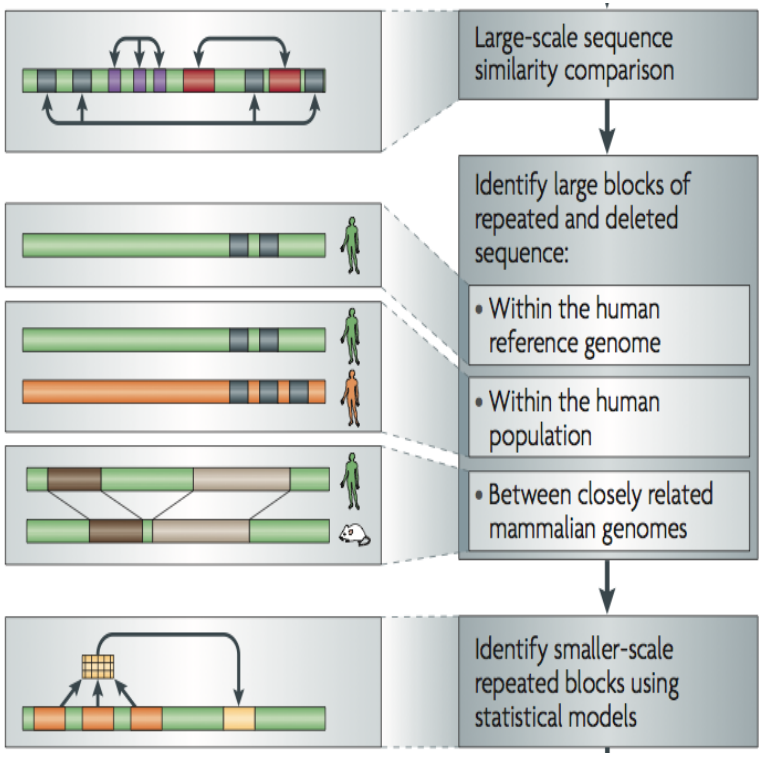
(last edit in spring ’20)

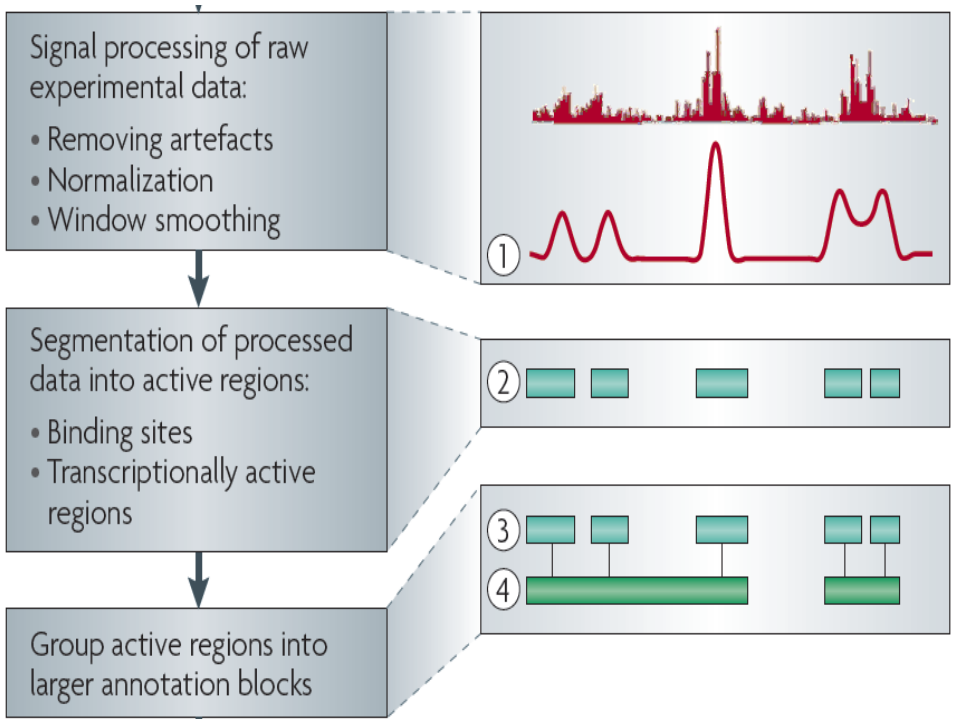
Non-coding Annotations: Overview

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

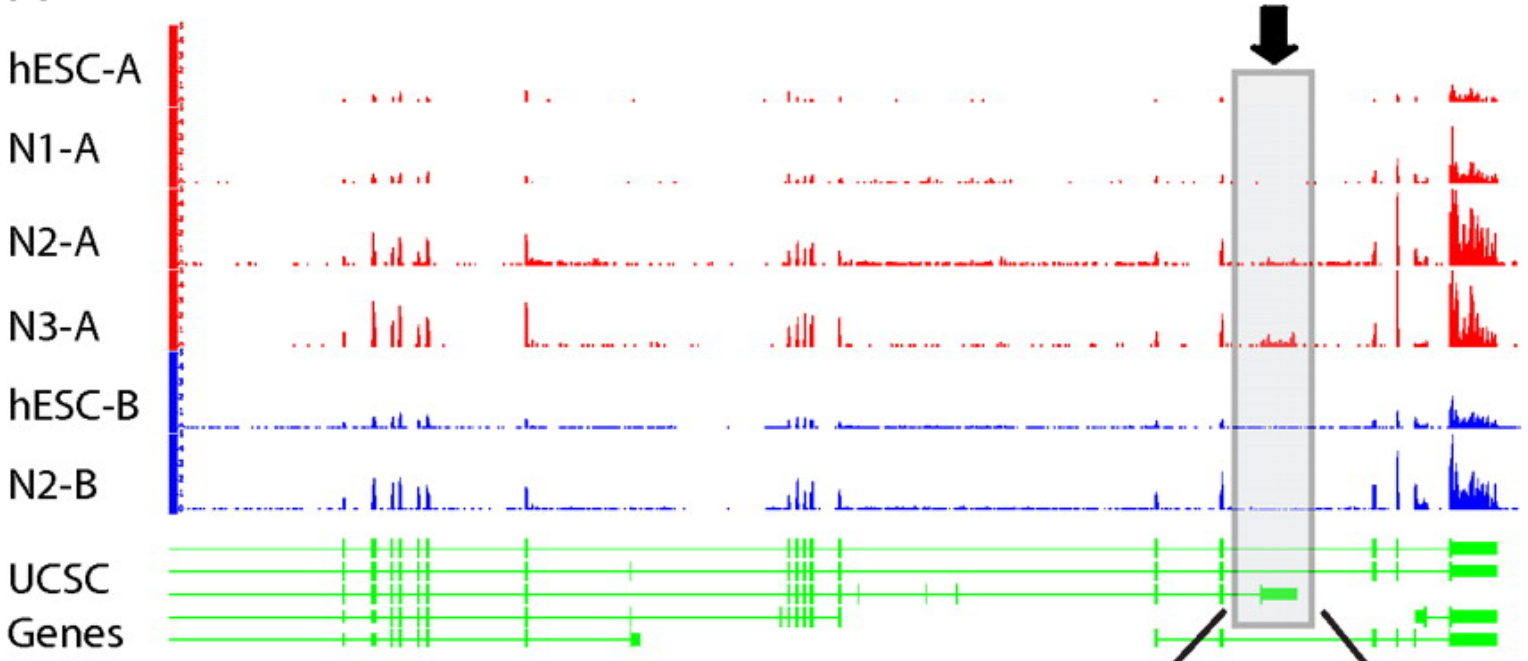
Sequence features, incl. **Conservation**

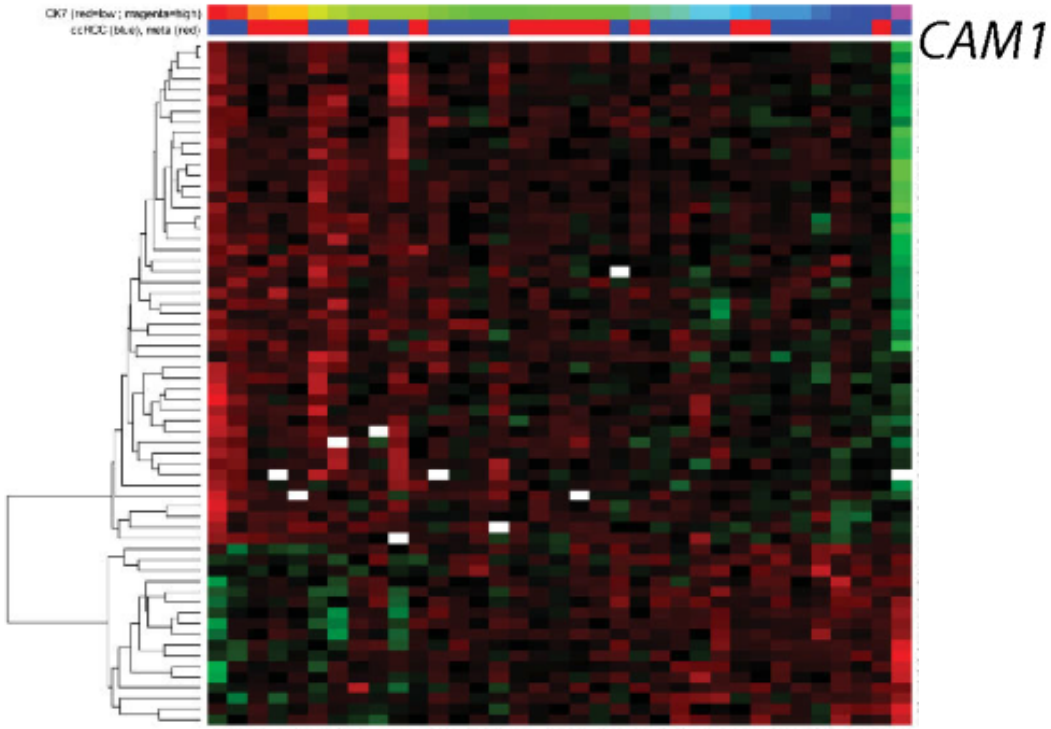
Functional Genomics
Chip-seq (Epigenome & seq. specific TF) and ncRNA & un-annotated transcription





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

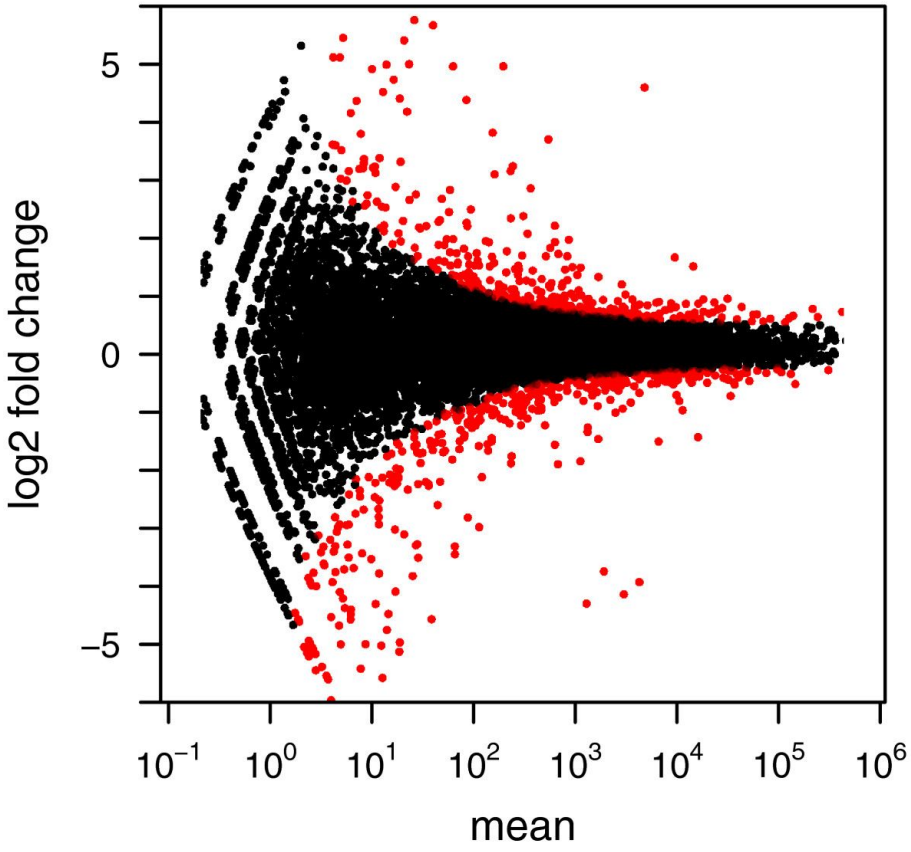




Information from RNA-seq: Avg. signal at exons & TARs (RPKMs)

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

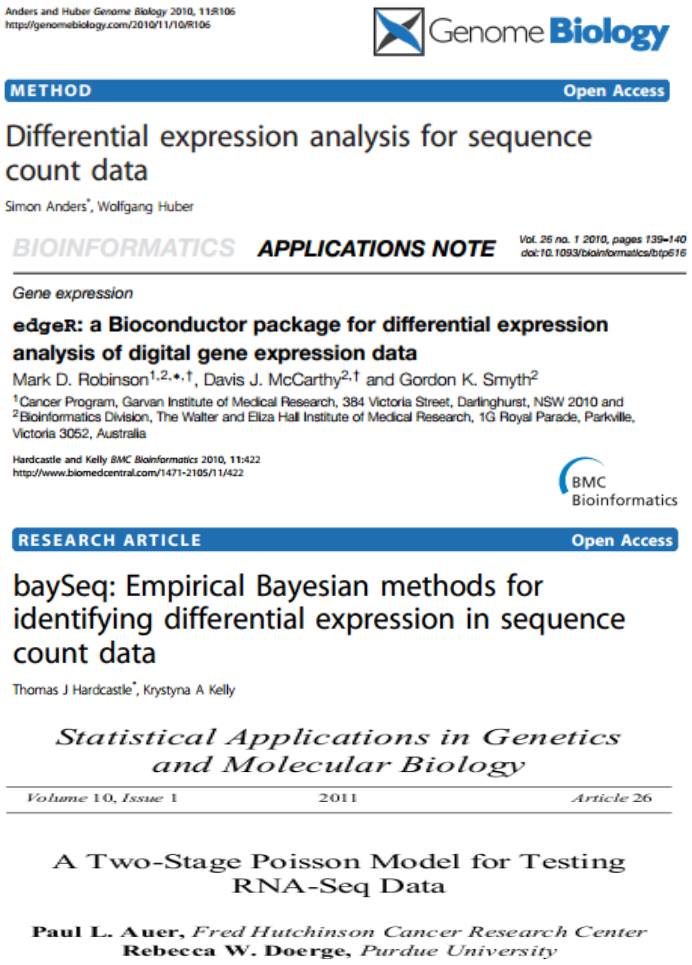
Differential expression analysis



Genome Biology, 2010 11:R106

Differential expression analysis: Count-based

1. **DESeq** -- based on negative binomial distribution
2. **edgeR** -- use an overdispersed Poisson model
3. **baySeq** -- use an empirical Bayes approach
4. **TSPM** -- use a two-stage poisson model



Anders and Huber *Genome Biology* 2010, 11:R106
http://genomebiology.com/2010/11/10/R106

Genome Biology

METHOD Open Access

Differential expression analysis for sequence count data

Simon Anders*, Wolfgang Huber

BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 1 2010, pages 139–140 doi:10.1093/bioinformatics/btp616

Gene expression

edgeR: a Bioconductor package for differential expression analysis of digital gene expression data

Mark D. Robinson[1,2,*,†], Davis J. McCarthy[2,†] and Gordon K. Smyth[2]

[1]Cancer Program, Garvan Institute of Medical Research, 384 Victoria Street, Darlinghurst, NSW 2010 and [2]Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, Victoria 3052, Australia

Hardcastle and Kelly *BMC Bioinformatics* 2010, 11:422
http://www.biomedcentral.com/1471-2105/11/422

BMC Bioinformatics

RESEARCH ARTICLE Open Access

baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data

Thomas J Hardcastle*, Krystyna A Kelly

Statistical Applications in Genetics and Molecular Biology

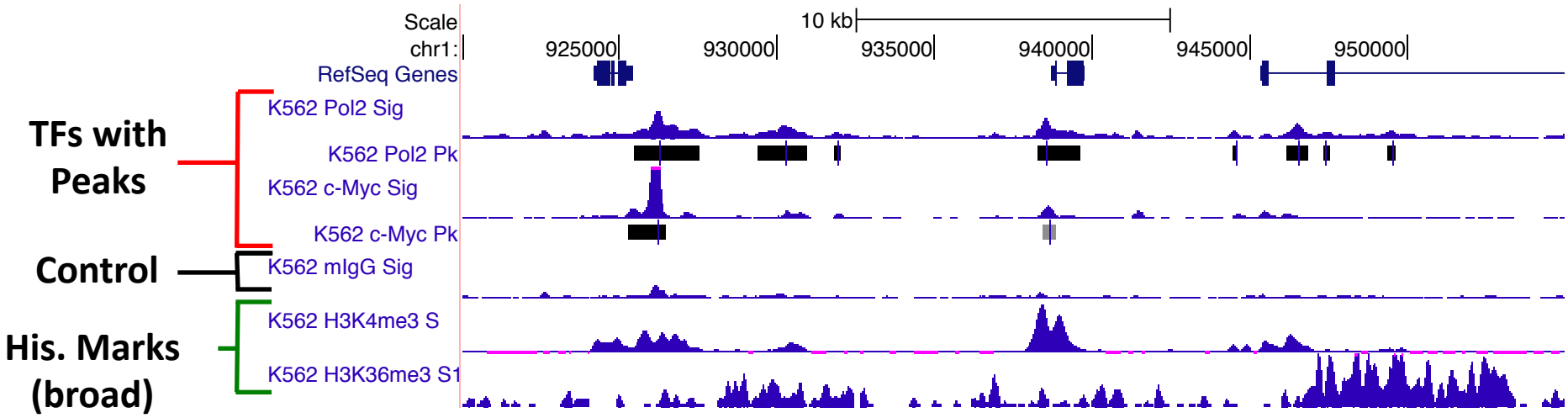
Volume 10, *Issue* 1 2011 *Article* 26

A Two-Stage Poisson Model for Testing RNA-Seq Data

Paul L. Auer, *Fred Hutchinson Cancer Research Center*
Rebecca W. Doerge, *Purdue University*

chip-seq

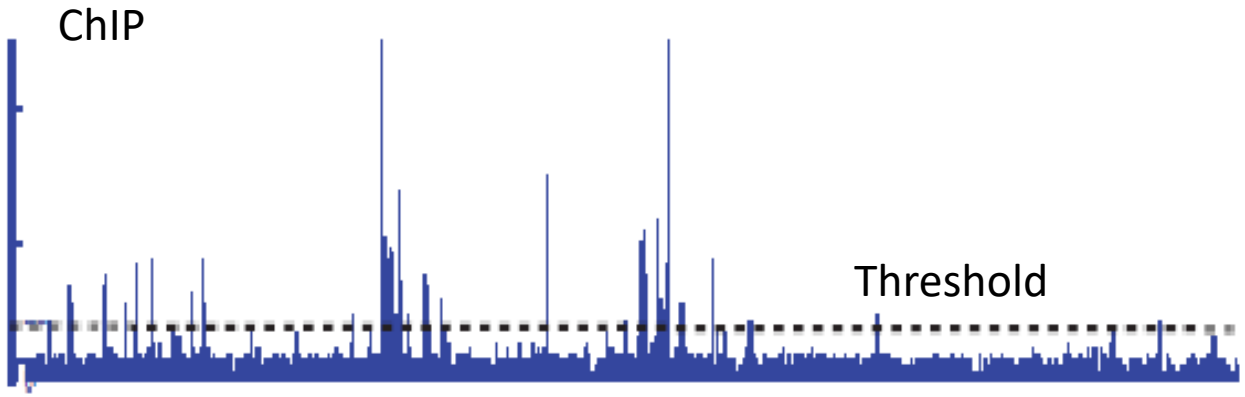
Information from Chip-seq



[*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 →

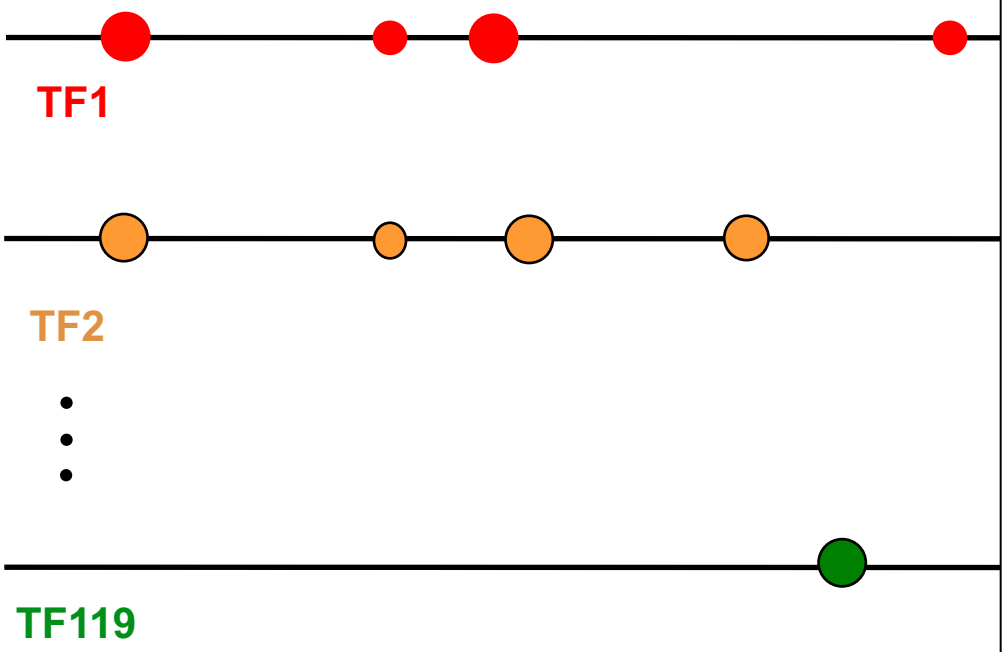
[Rozowsky et al. ('09) *Nat Biotech*]

Data Flow: Chip-seq expts. to co-associating peaks

119 TFs from 458 ChIP-Seq experiments (2 Tb tot.)

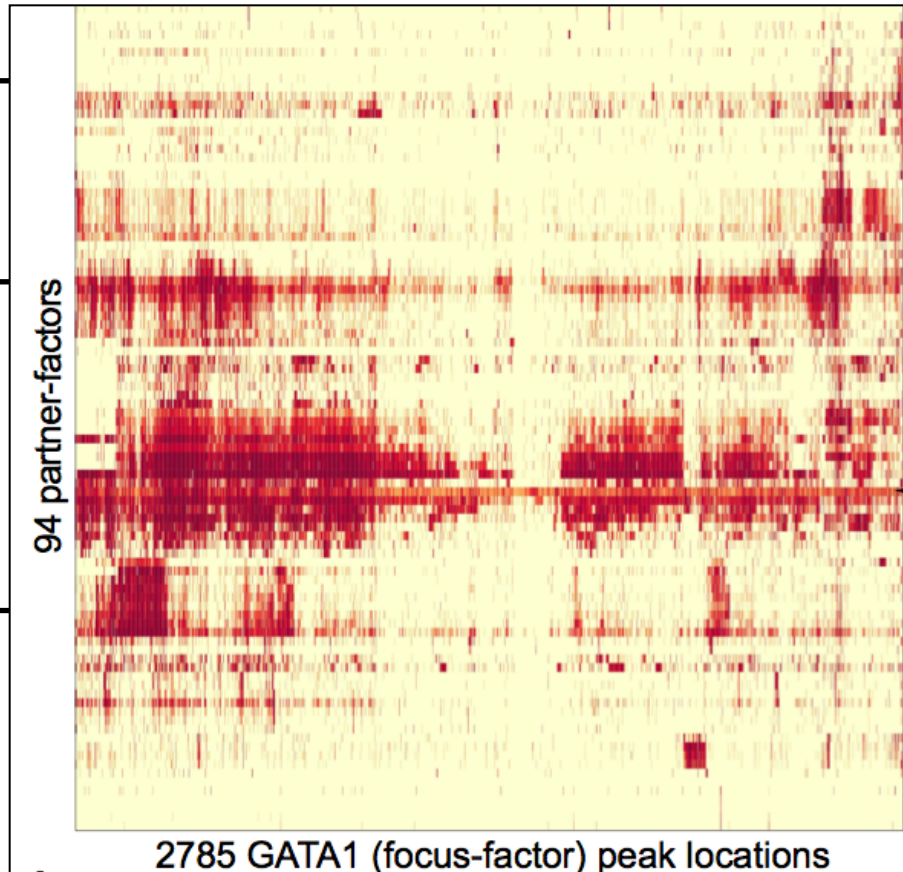
Signal Tracks

7M Peaks from Uniform Peak Calling



- Mostly in Tier 1 cell lines
  - K562, GM12878, H1h-ESC…
- Matching RNA-Seq data in all cell-lines

- SPP & PeakSeq
- thresholding w. IDR (replicas)



[Gerstein et al. Nature (in press, '12)]

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

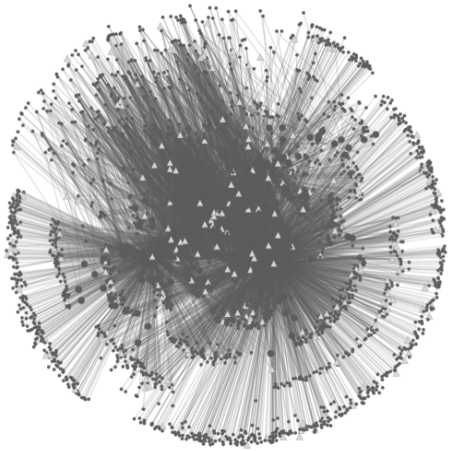
TF

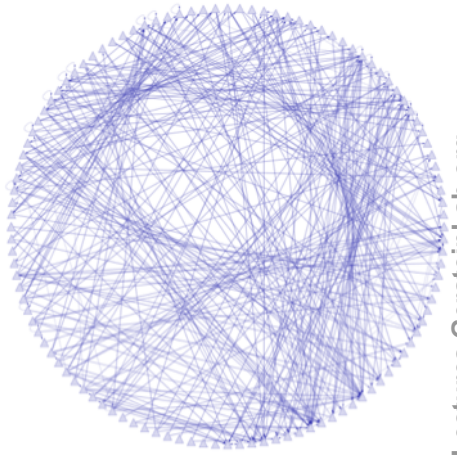
Potential Distal Edge

TF

Strong Proximal Edge

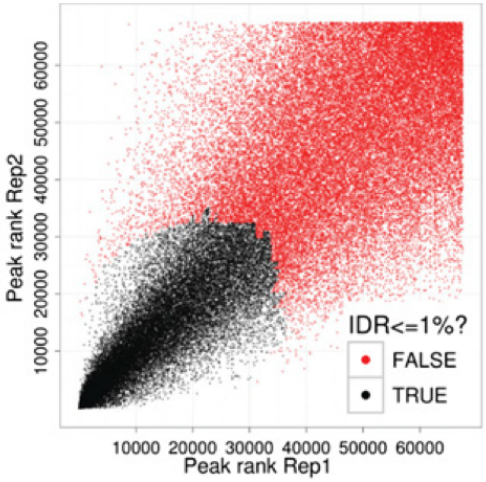
[Cheng et al., *Bioinfo.* ('11);
Gerstein et al. *Nature* (in press, '12) ;
Yip et al., *GenomeBiology* (in press, '12)]

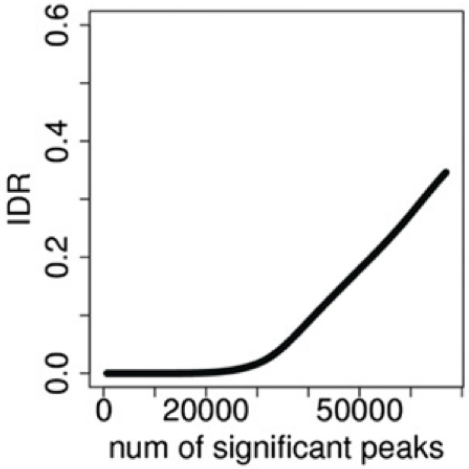




The irreproducible discovery rate (IDR)

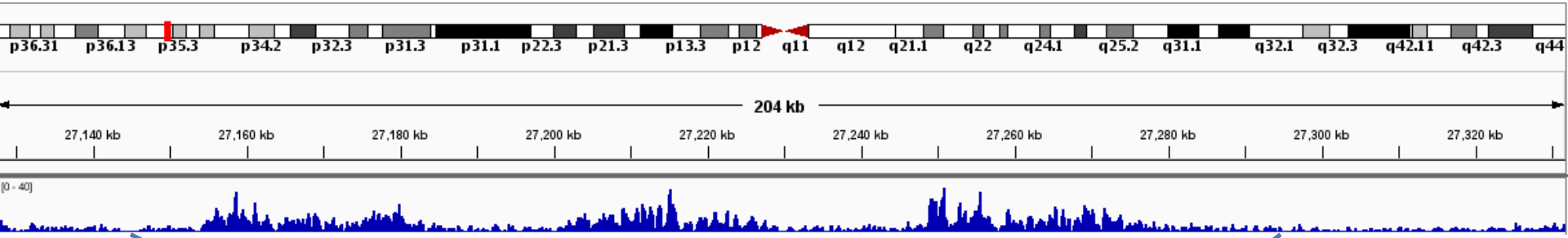
- Unified approach to measure the reproducibility of findings identified from replicate high-throughput experiments.
- Idea : call peaks with low cutoff and classify peaks as reproducible or not (bivariate rank distributions) based on overlap of ranked peaks (consistency)

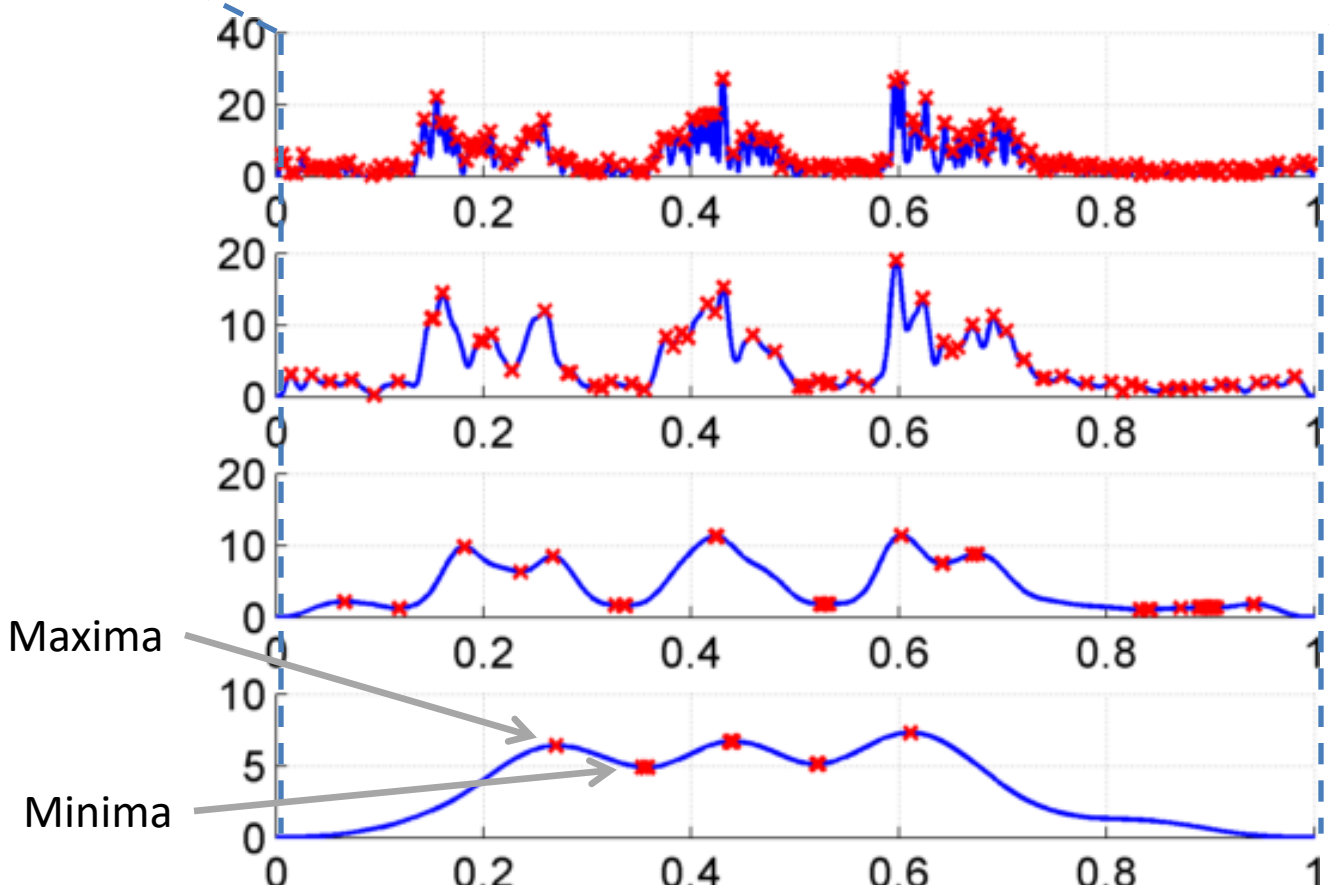




Genome Research 2012, 22:1813-1831

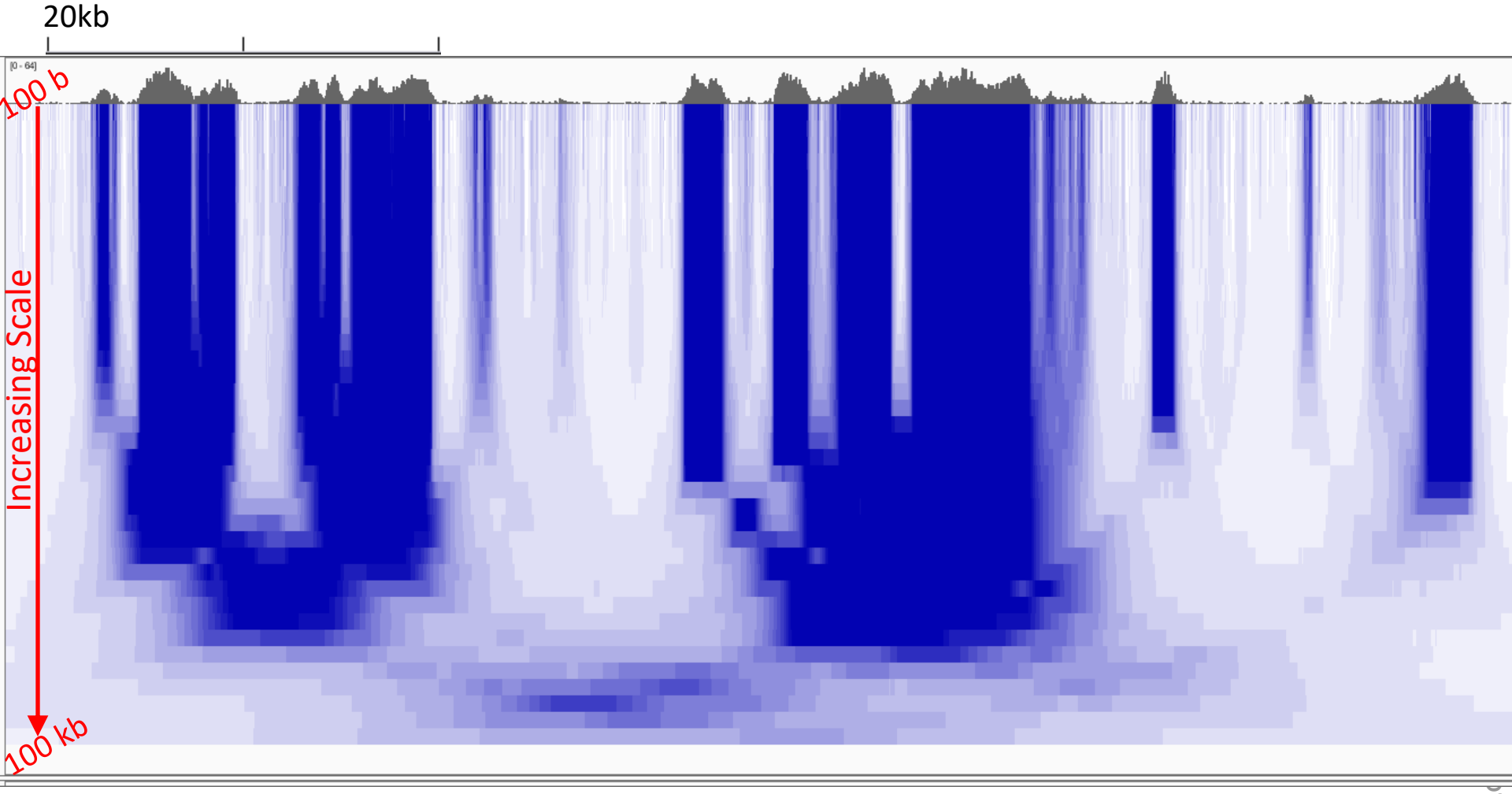
Multiscale Analysis, Minima/Maxima based Coarse Segmentation





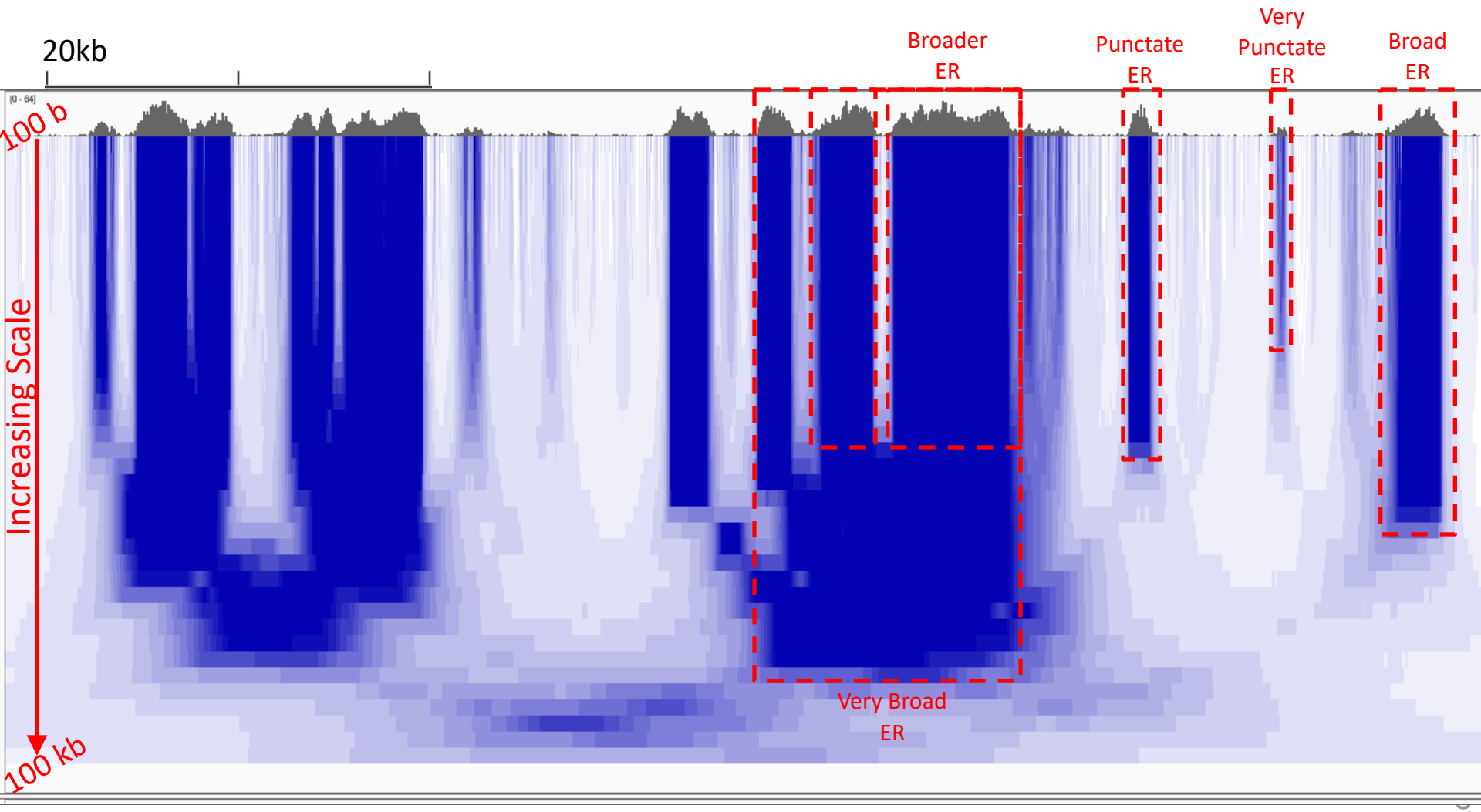
Harmanci et al, Genome Biology 2014, MUSIC.gersteinlab.org

Multiscale Decomposition



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

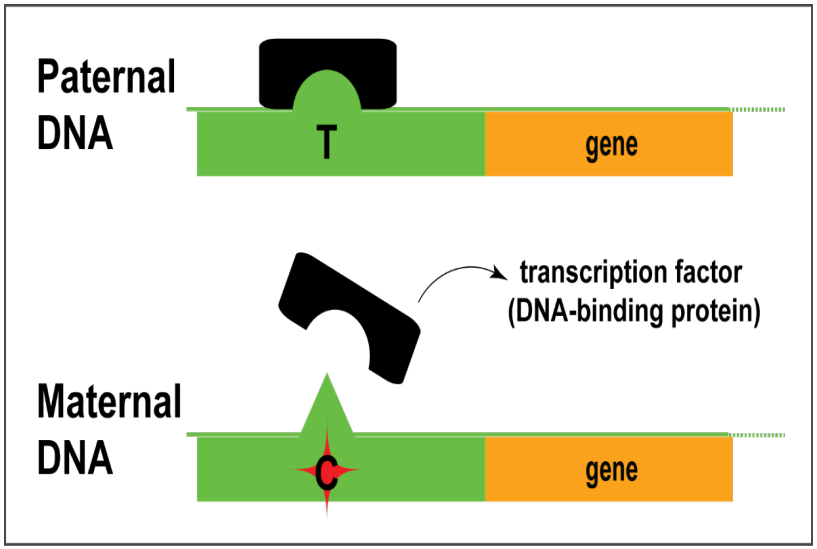
Multiscale Decomposition



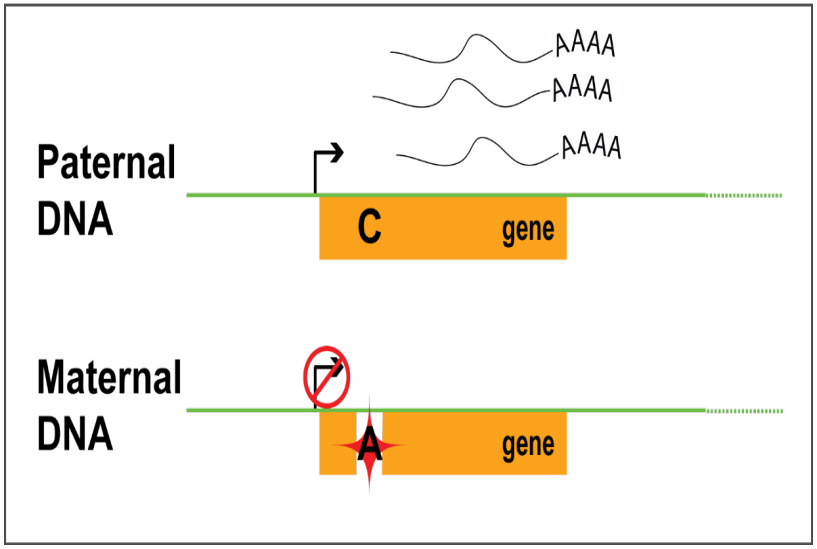
[Harmanci *et al, Genome Biol.* ('14)]

ASB/ASE & eQTL

Allele-specific binding and expression



Genomic variants affecting allele-specific behavior e.g. allele-specific binding (ASB)



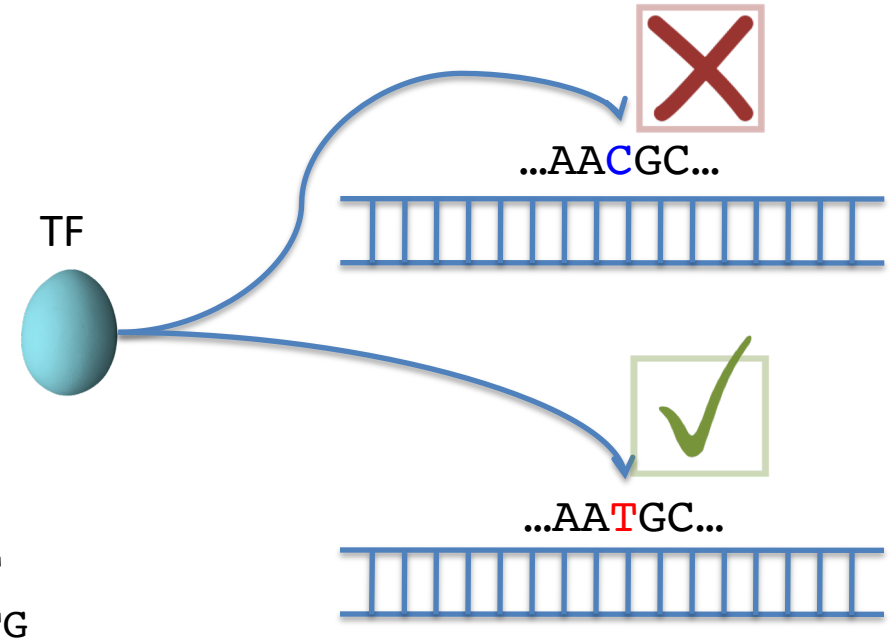
e.g. allele-specific expression (ASE)

Inferring Allele Specific Binding/Expression using Sequence Reads

RNA/ChIP-Seq Reads

```
ACTTTGATAGCGTCAATG
 CTTTGATAGCGTCAATGC
 CTTTGATAGCGTCAACGC
   TTGACAGCGTCAATGCAC
    TGATAGCGTCAATGCACG
      ATAGCGTCAATGCACGTC
       TAGCGTCAATGCACGTCG
          CGTCAACGCACGTCGGGA
           GTCAATGCACGTCGAGAG
             CAATGCACGTCGGGAGTT
              AATGCACGTCGGGAGTTG
                TGCACGTTGGGAGTTGGC
```

```
10 x T
 2 x C
```



Haplotypes with a Heterozygous Polymorphism

Interplay of the annotation and individual sequence variants

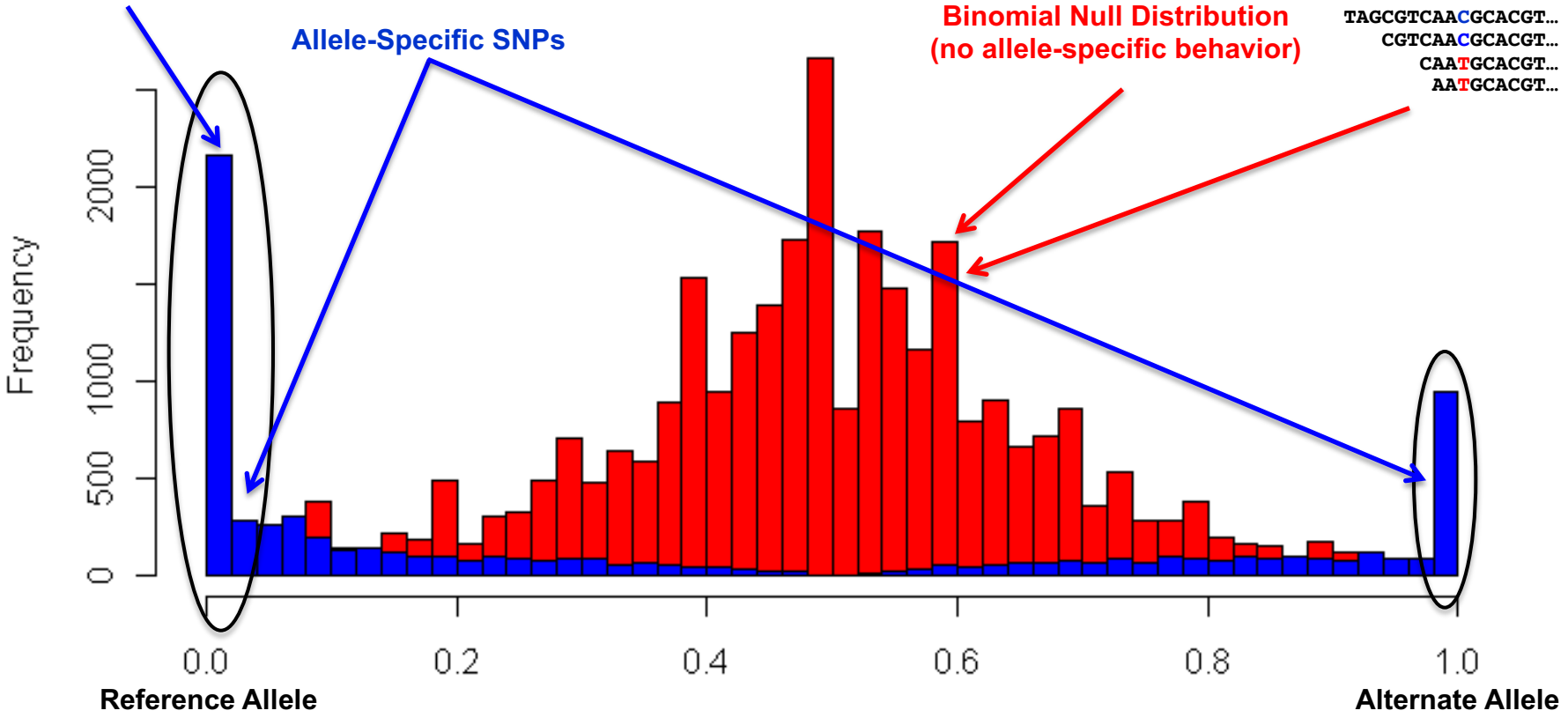
Many Technical Issues in Determining ASE/ASB: Reference Bias (naïve alignment against reference)

ASE/ASB Example:

…GTCAATGCAC
…GTCAATGCACG
…GTCAATGCACGTC
…GTCAATGCACGTCG
…GTCAACGCACGTCGGGA
GTCAATGCACGTCGAGAG
CAATGCACGTCGGGAGTT
AATGCACGTCGGGAGTTG

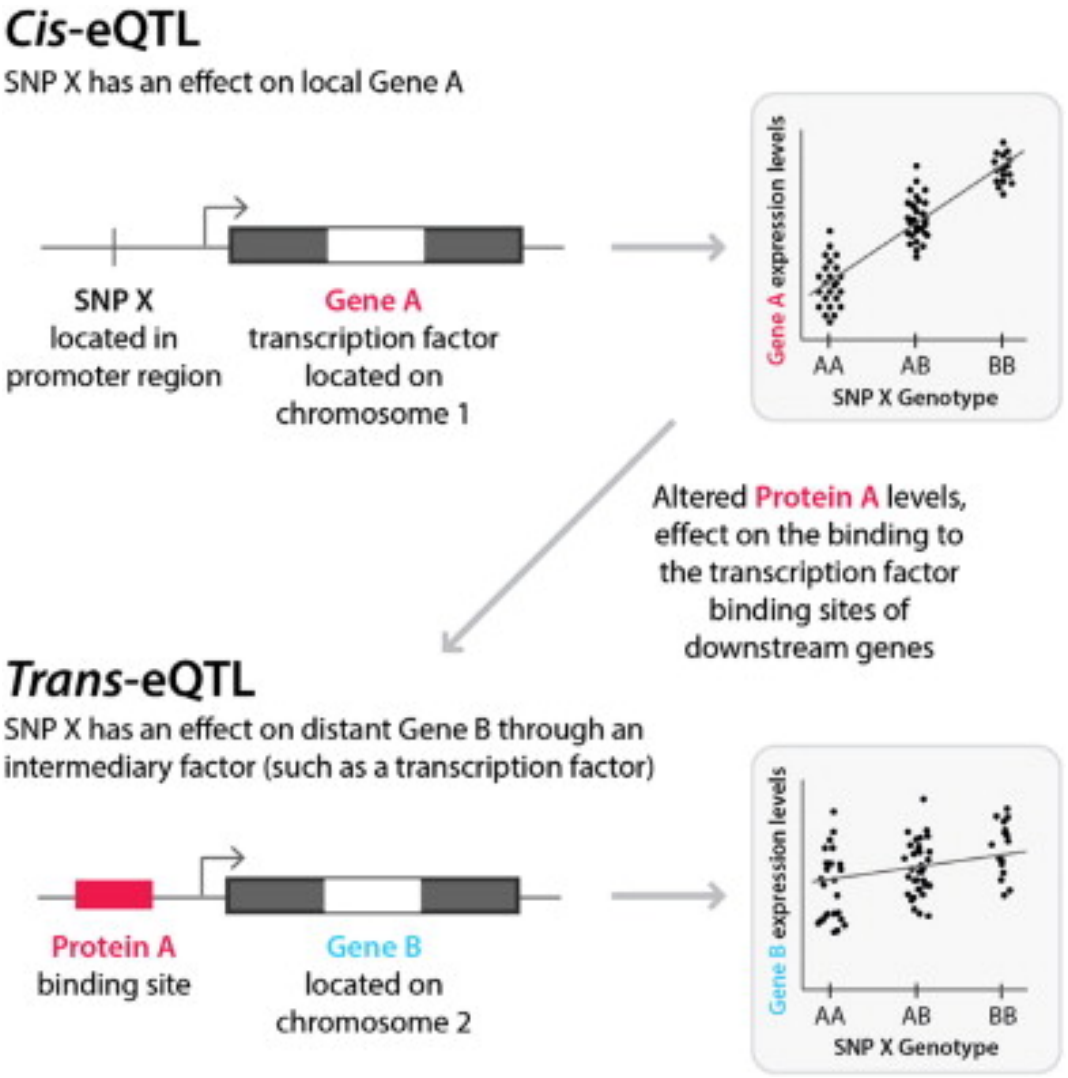
Null Example:

ACTTTGATAGCGTCAATG
CTTTGATAGCGTCAACGC
TTGACAGCGTCAATGCAC
ATAGCGTCAATGCACGT…
TAGCGTCAACGCACGT…
CGTCAACGCACGT…
CAATGCACGT…
AATGCACGT…

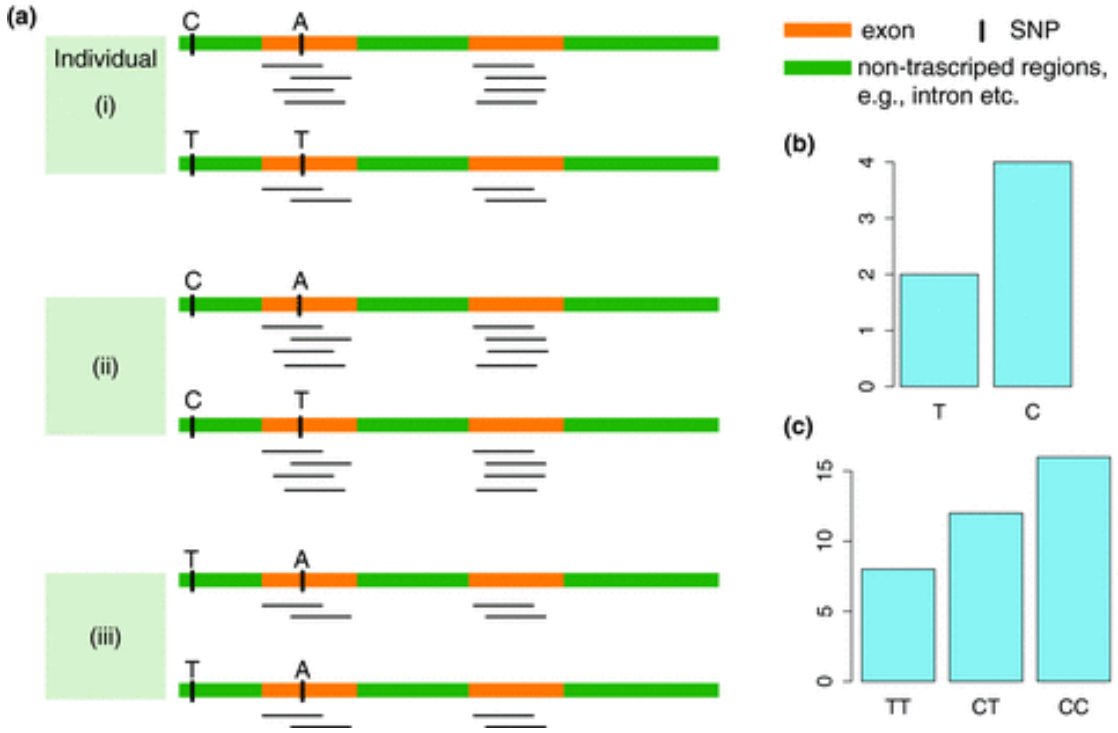


[Rozowsky et al., MSB ('11)]

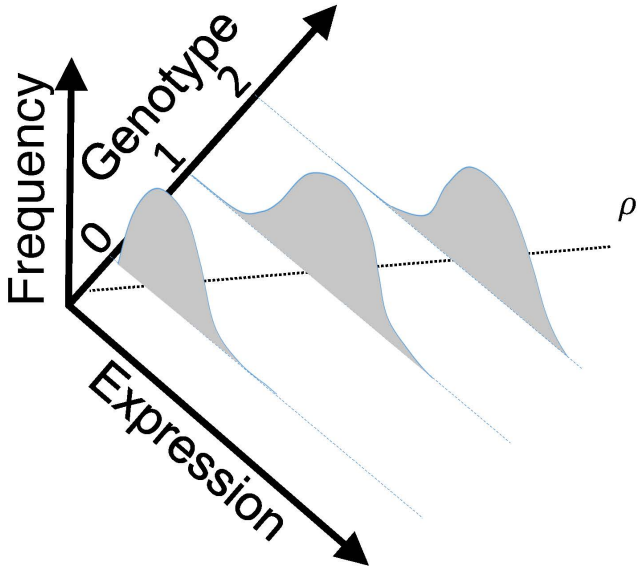
Expression quantitative trait



Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease 2014, 10:1896-1902



[*Biometrics 68(1) 1–11*]



eQTL Mapping Using RNA-Seq Data

- eQTLs are genomic loci that contribute to variation in mRNA expression levels
- eQTLs provide insights on transcription regulation, and the molecular basis of phenotypic outcomes
- eQTL mapping can be done with RNA-Seq data

Hi-C

3D organization of genome

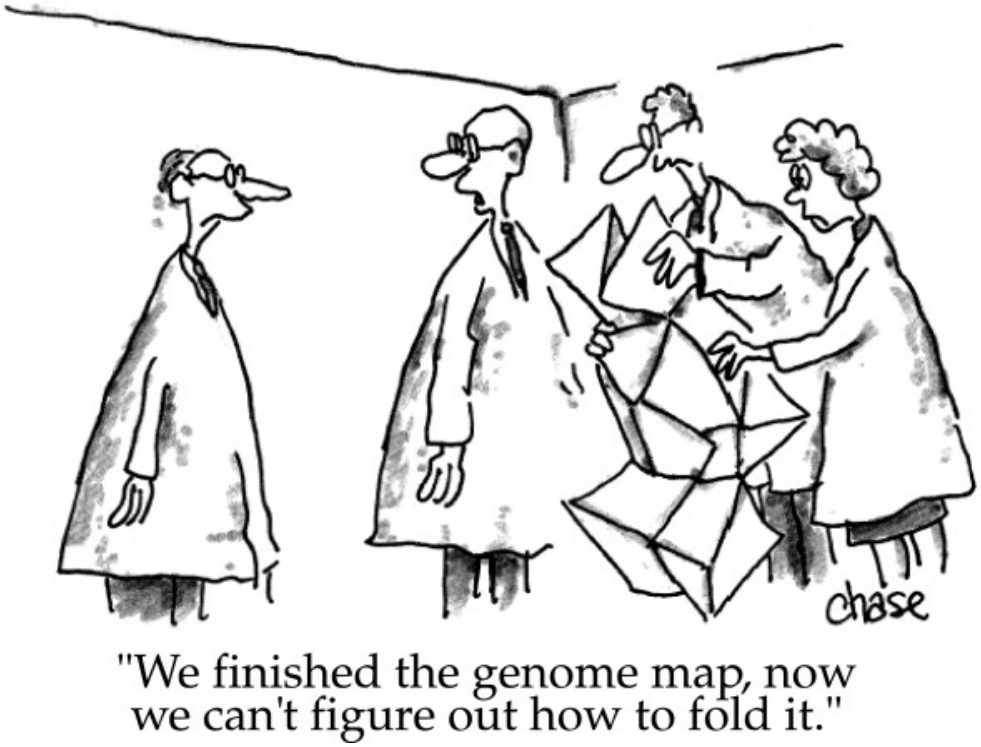


image credit: Iyer et al. BMC Biophysics 2011, cartoonist John Chase

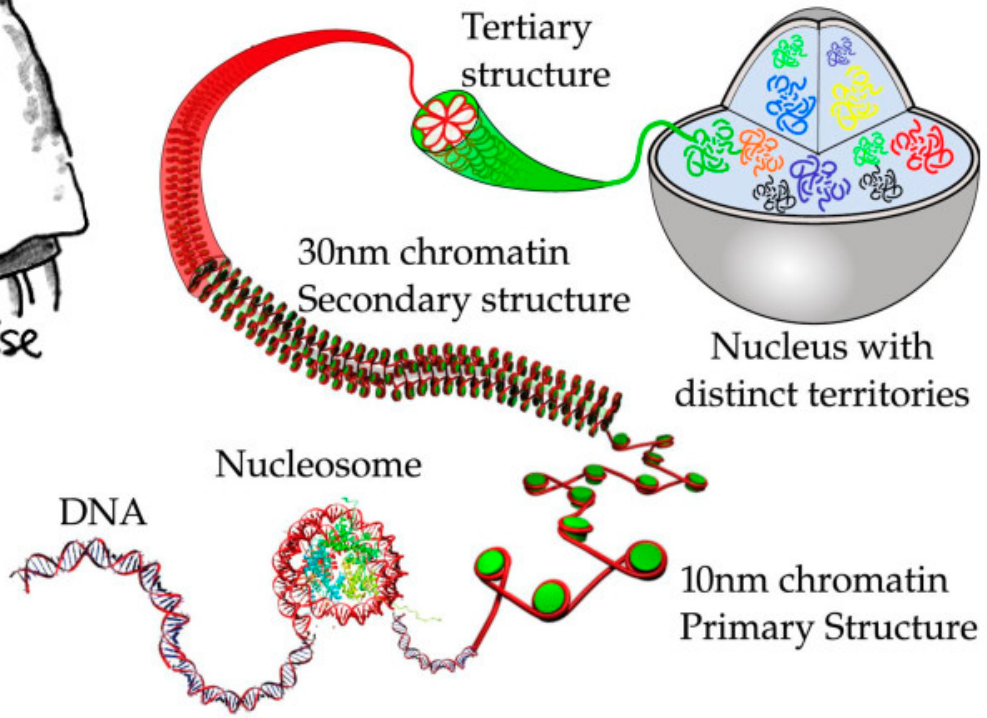
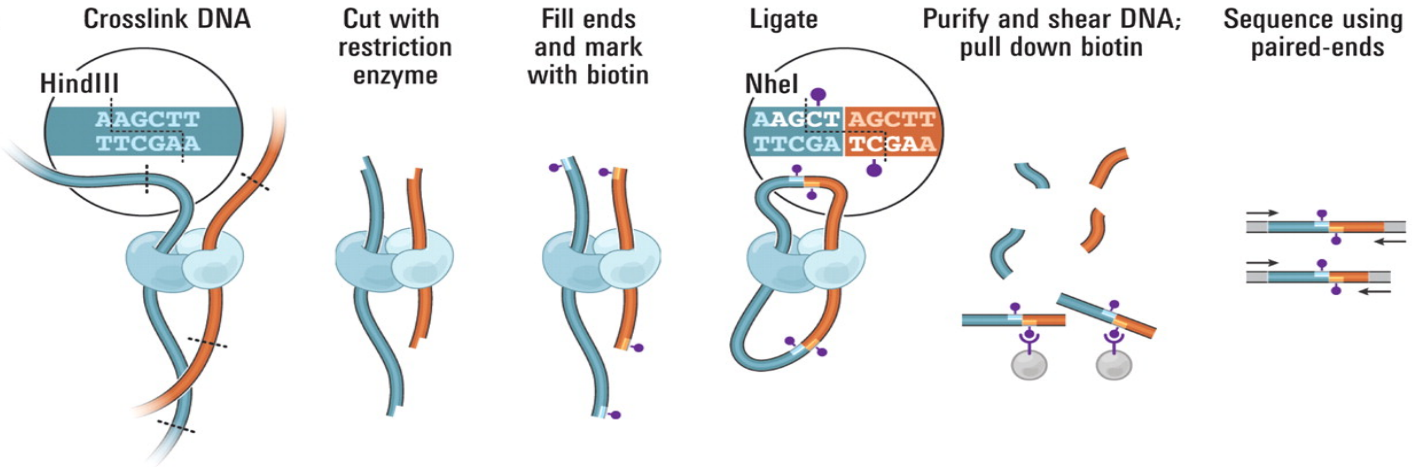
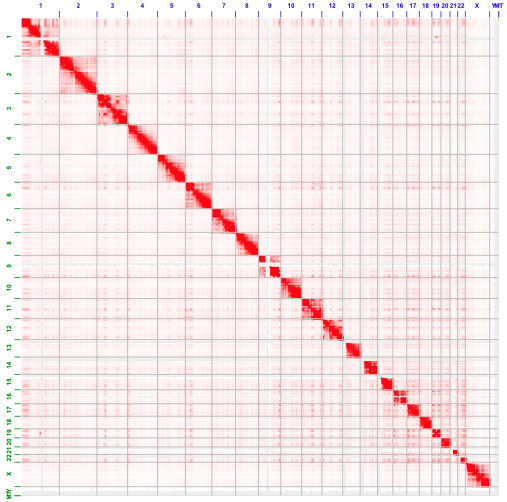


image credit: Iyer et al. BMC Biophysics 2011

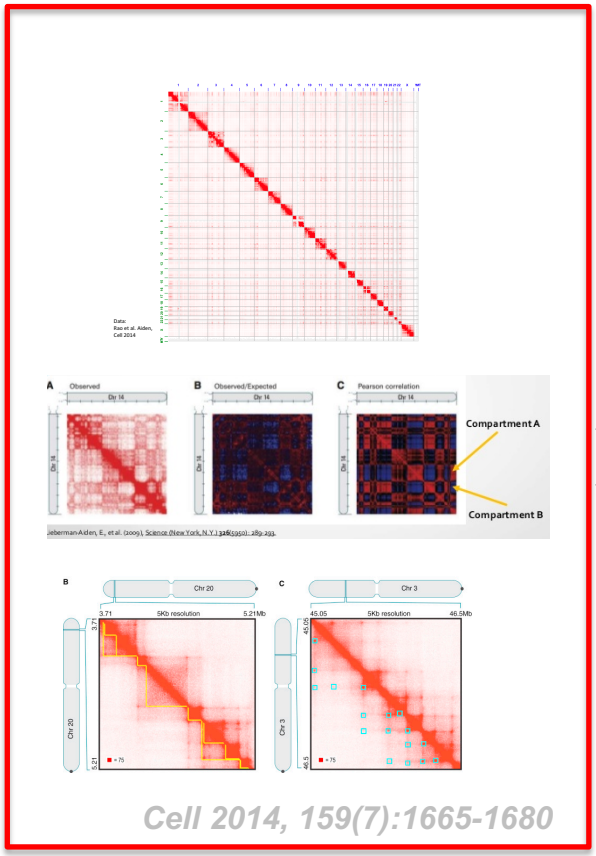
Hi-C contact map

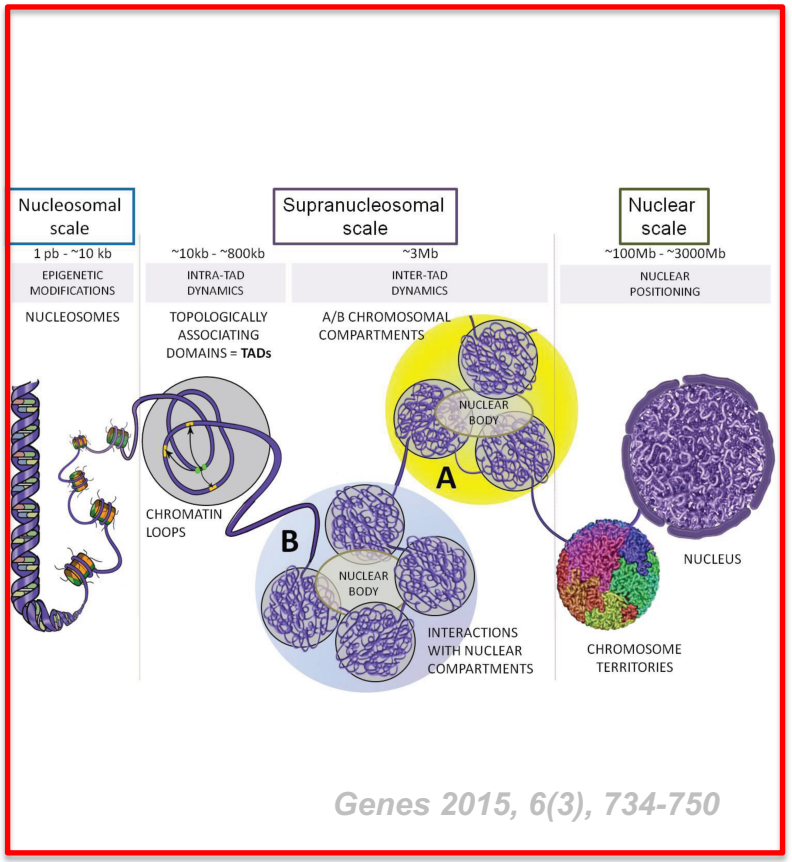




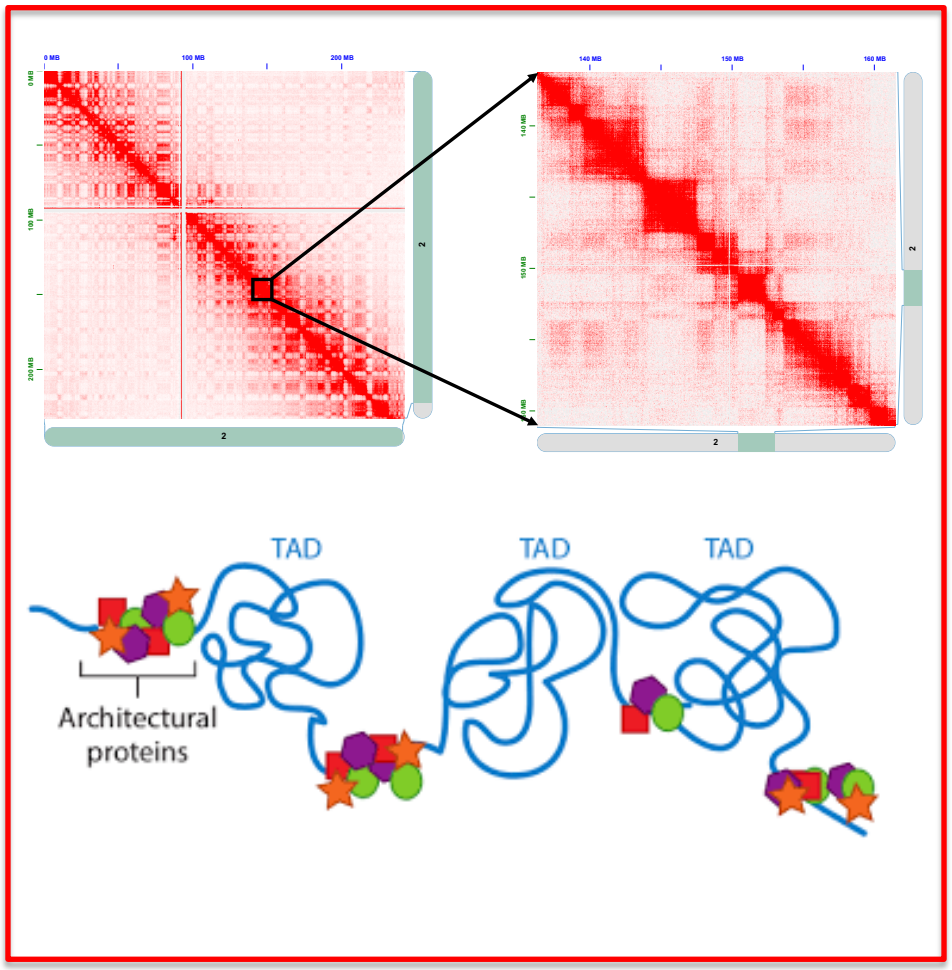
Science 2009, 5950: 289-293

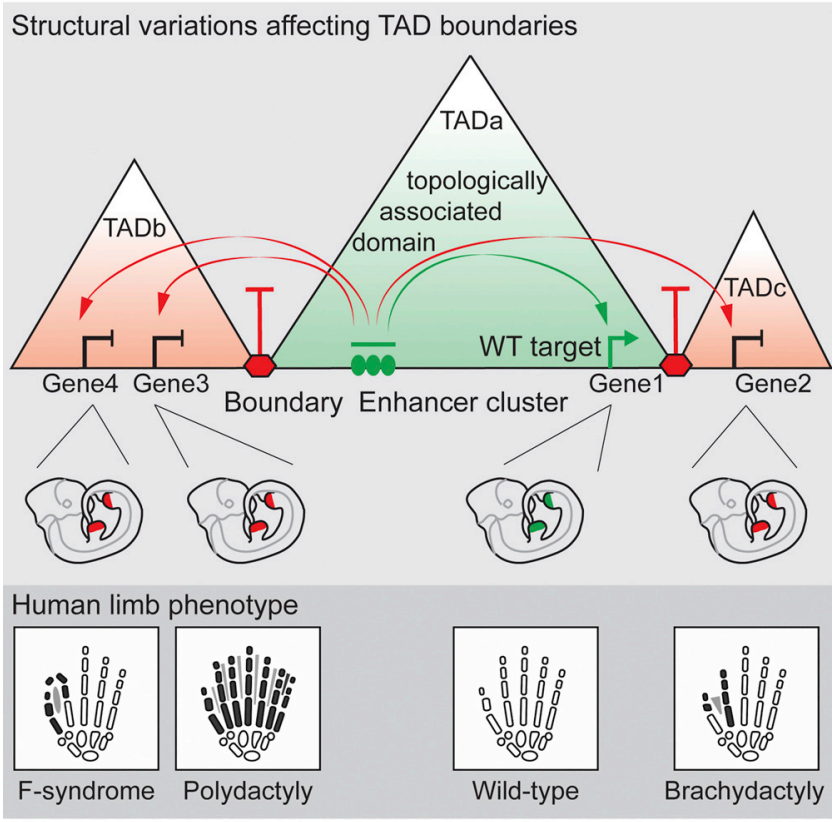
Hi-C contact map and Genome architecture





Topologically Associating Domain

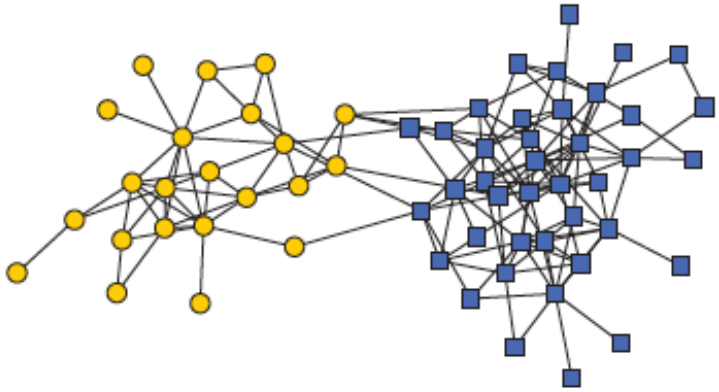




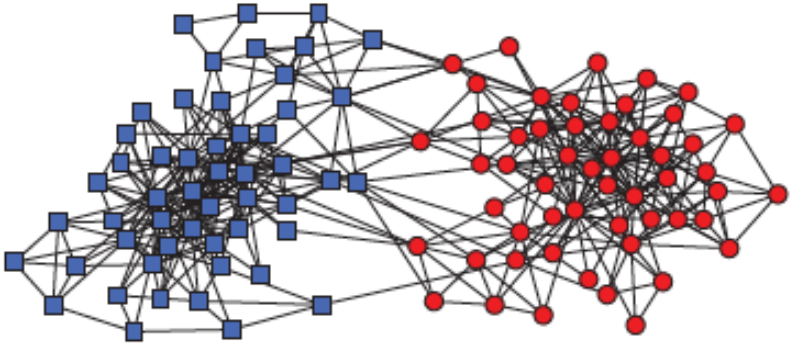
Cell 2015, 161:1012-1025

Modularity

Network modularity



Dolphin social network



Political books

Newman Phy. Rev. E 2013

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

adjacency matrix

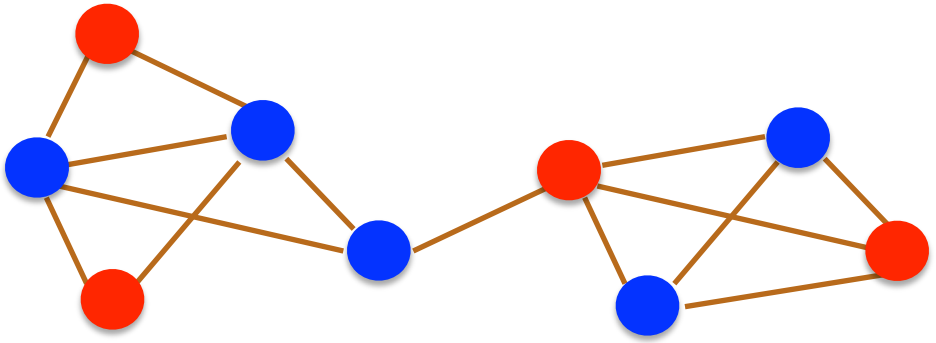
degree of node i

number of edges

expected number of edges between i and j

whether or not i, j are in the same module

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

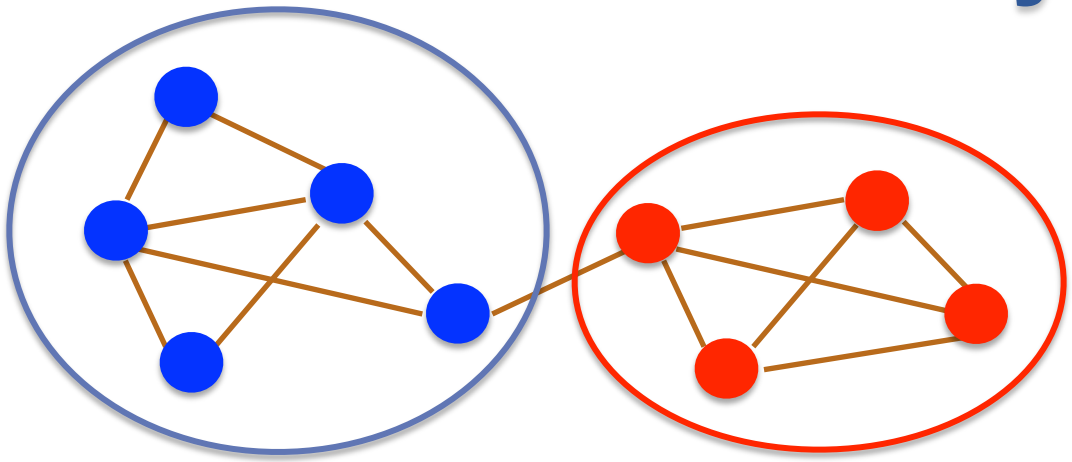
degree of node i

number of edges

expected number of edges between i and j

whether or not i, j are in the same module

Network modularity



$$Q = Q_{max}$$

Optimization problem for sim. annealing

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

adjacency matrix

degree of node i

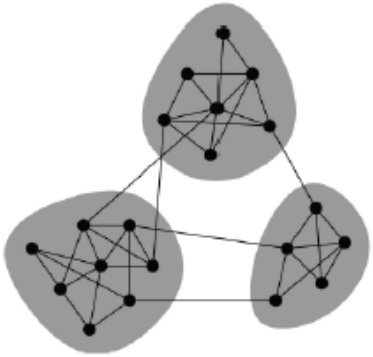
number of edges

expected number of edges between i and j

whether or not i, j are in the same module

TAD Finding

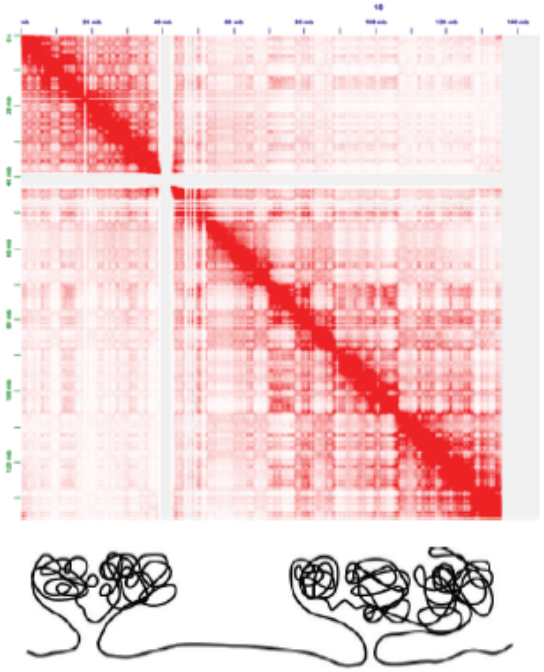
Identifying TADs in multiple resolutions



Modularity maximization

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

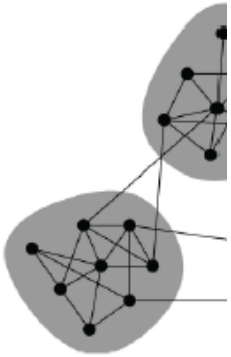
network	contact map
node	chromosome bin
edge	Hi-C contact
# of connections	coverage
module	domain



schematic adapted from ref. [2]

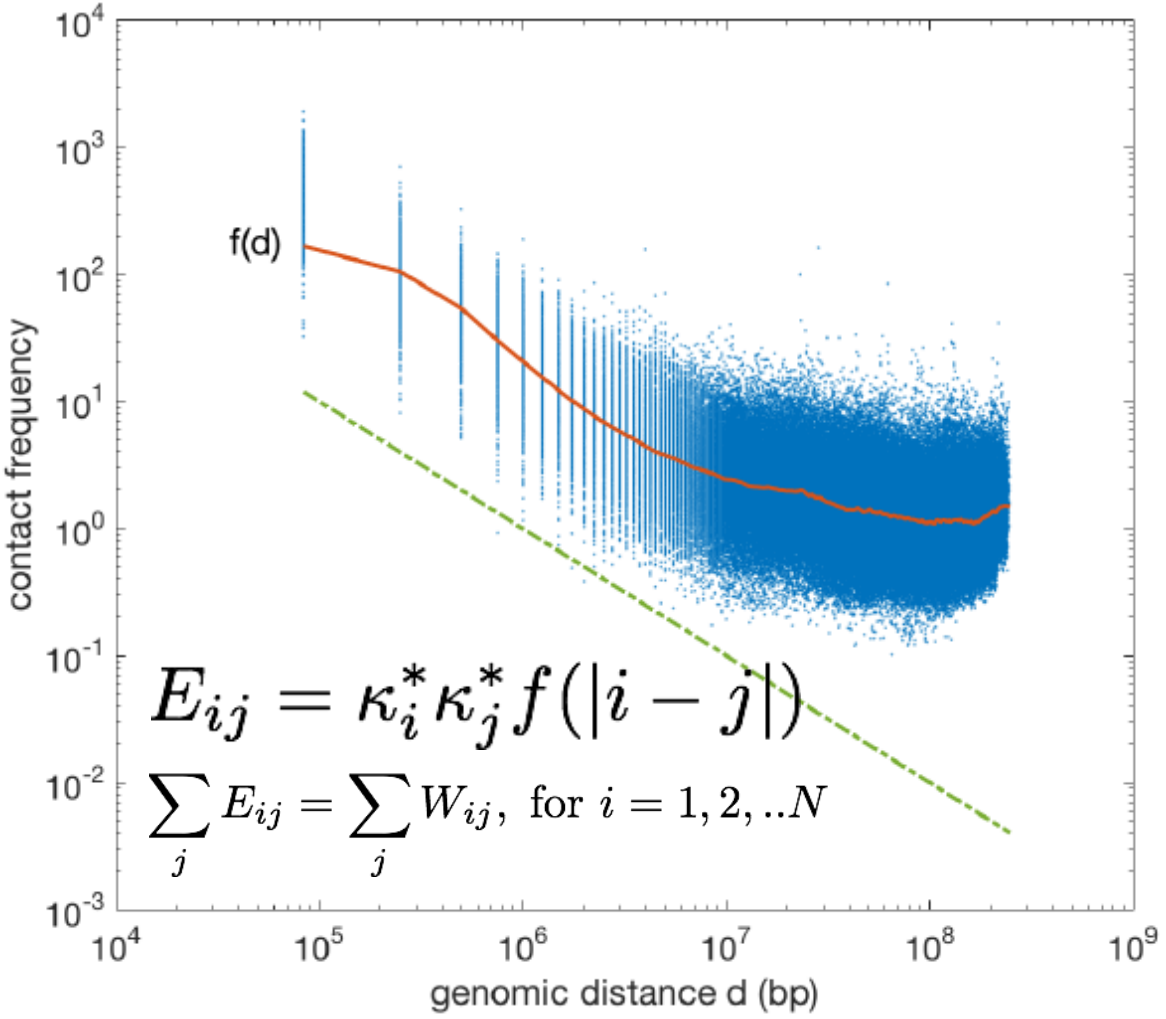
[Yan et al., *PLOS Comp. Bio.* (in revision, ‘17); bioRxiv 097345]

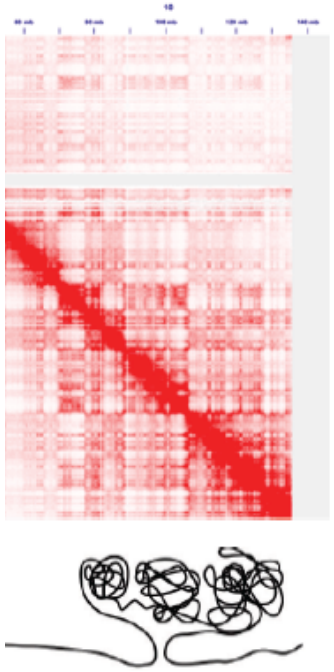
Identifying TADs in multiple resolutions



Modularity m

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



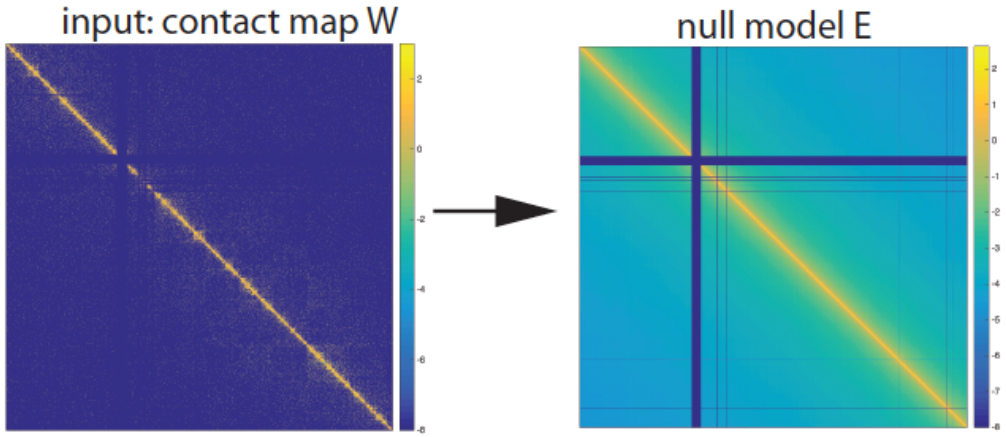


: adapted from ref. [2]

[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