3000788 Intro to Comp Molec Biol

Week 9: Proteomics

Fall 2024

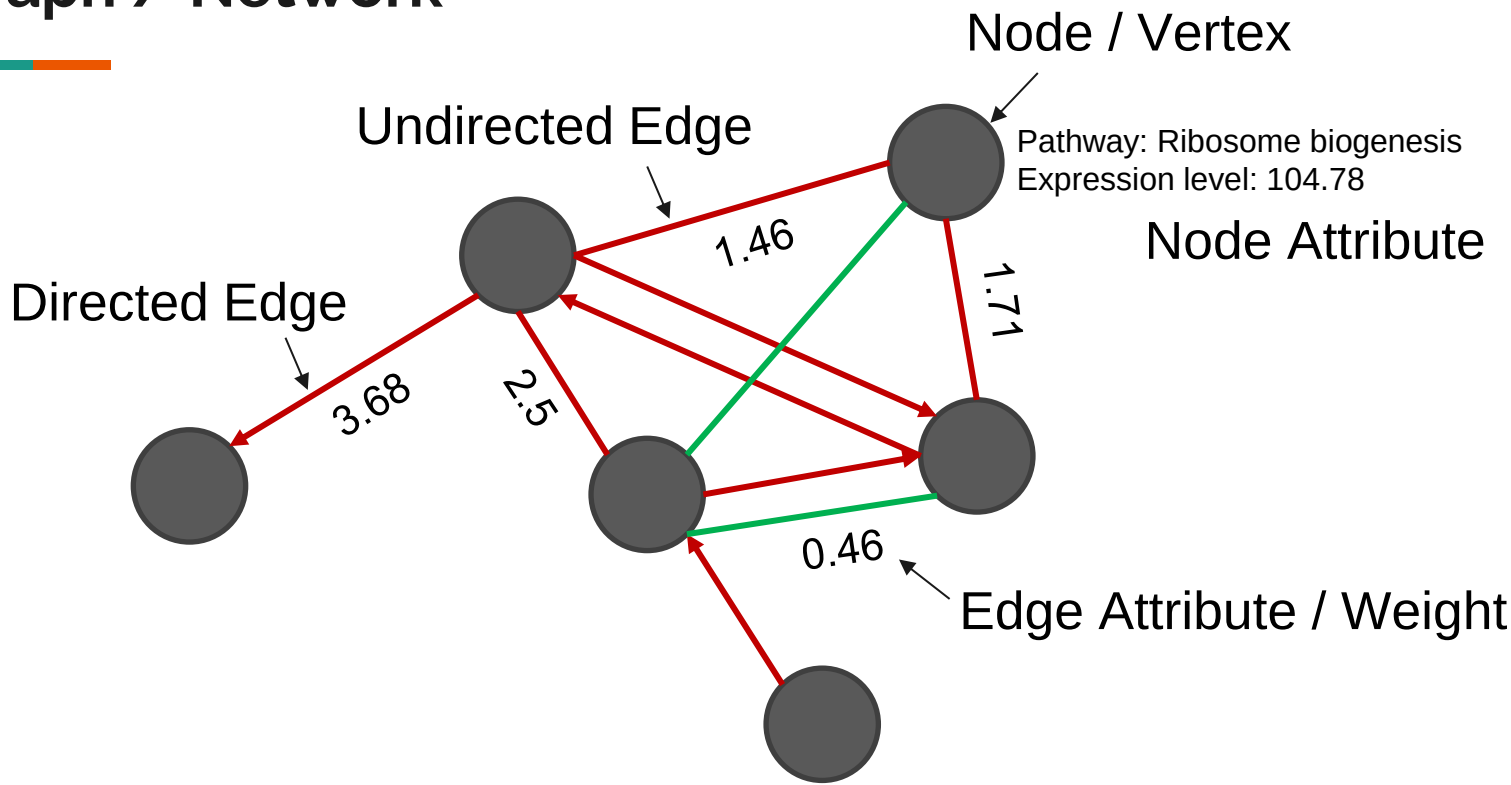
Sira Sriswasdi, PhD

- Research Affairs
- Center of Excellence in Computational Molecular Biology (CMB)
- Center for Artificial Intelligence in Medicine (CU-AIM)

Part 1: Biological Networks

- Network = nodes + edges
- Network capture interactions and communication
- Perturbation of a gene can propagate to other genes through interactions
- Network motif and module

Graph / Network



- Connection & relationship between entities

Real-world networks

- Computer network
- City-street
- Internet webpages
- Co-authorship
- Friendship
- River & sewage

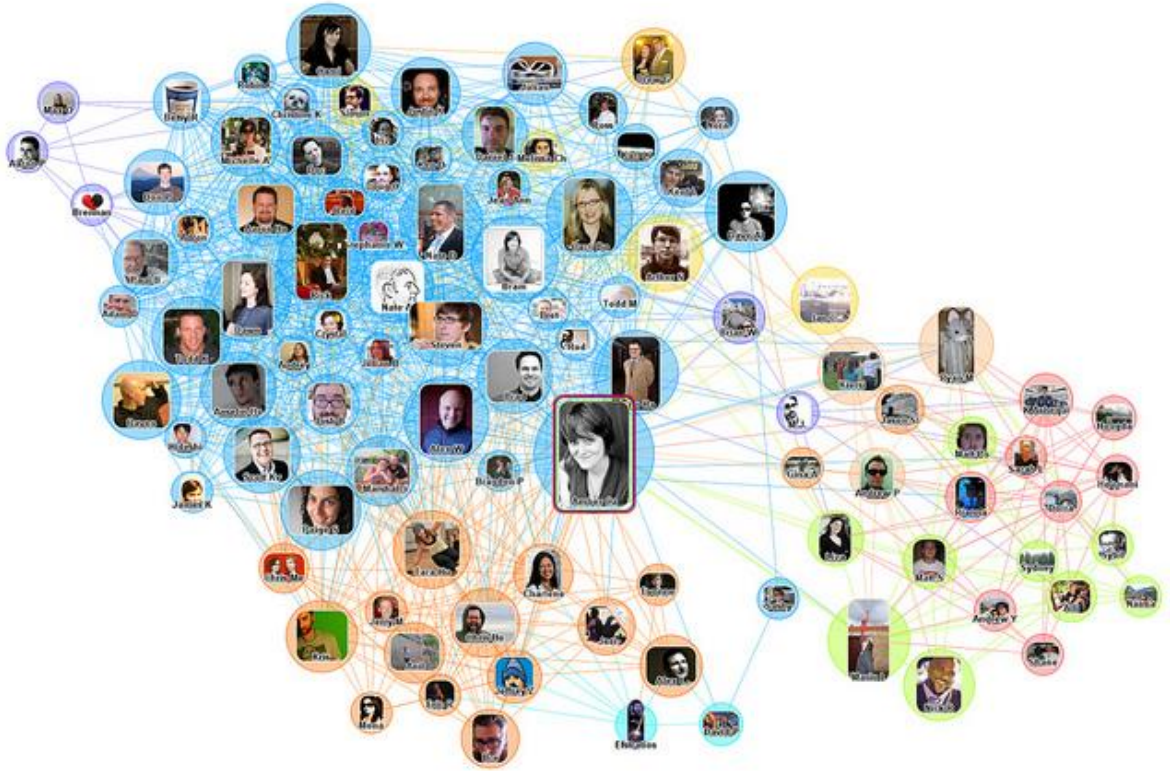
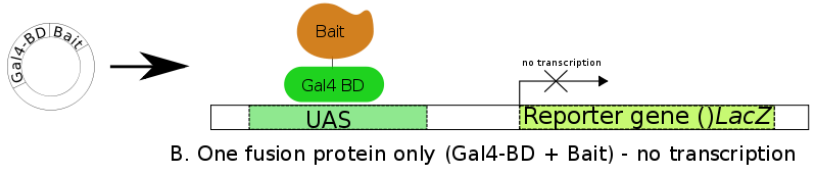


Image from https://www.flickr.com/photos/caseorganic/4935751455

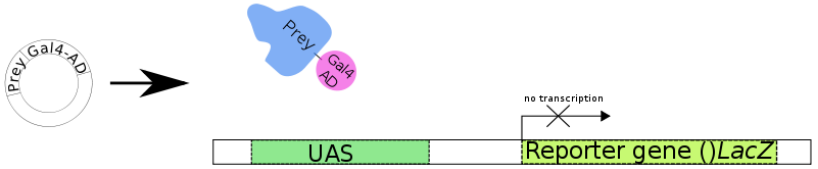
Biological networks

Networks from omics data

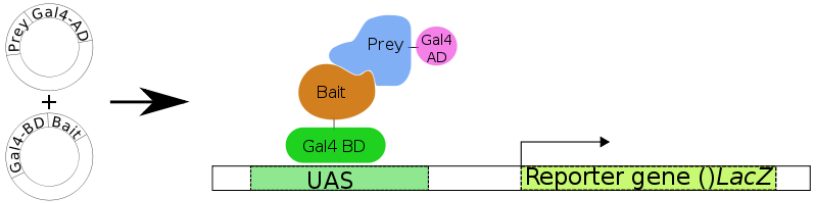
- Yeast-2-hybrid → protein-protein interaction networks
- Immunoprecipitation-MS → protein complex
- ChIP-seq → TF-gene regulatory network
- RNA-seq → gene co-expression network



B. One fusion protein only (Gal4-BD + Bait) - no transcription



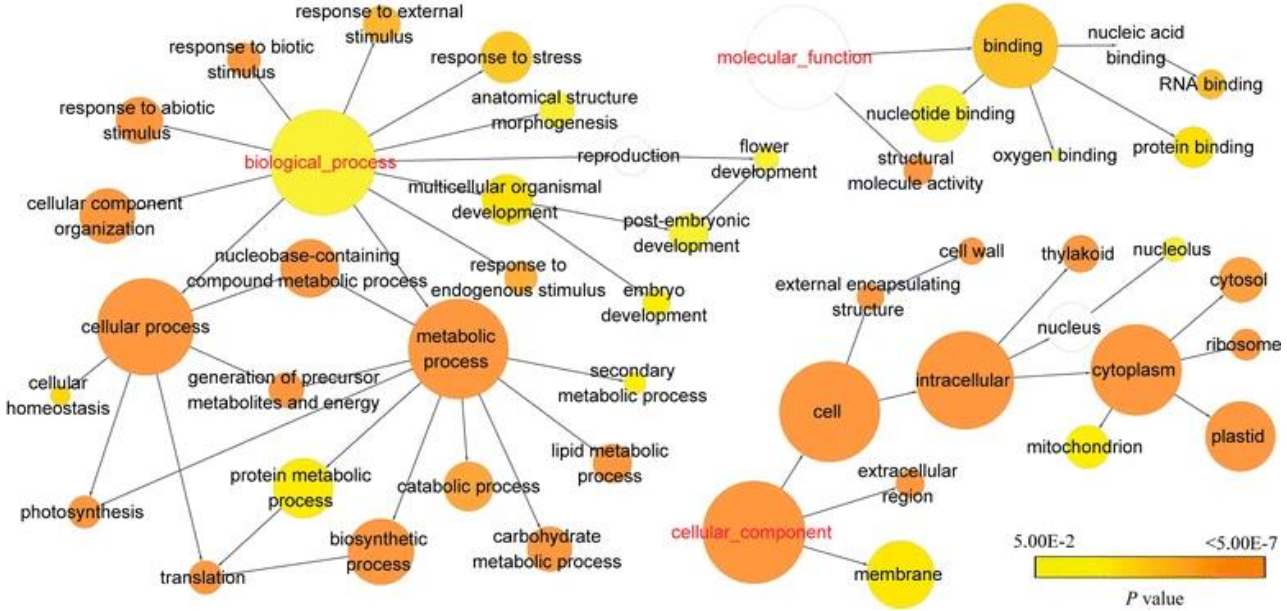
C. One fusion protein only (Gal4-AD + Prey) - no transcription



D. Two fusion proteins with interacting Bait and Prey

Source: Wikipedia.com

Gene ontology network



Gao, B. et al. BMC Genomics 16:416 (2015)

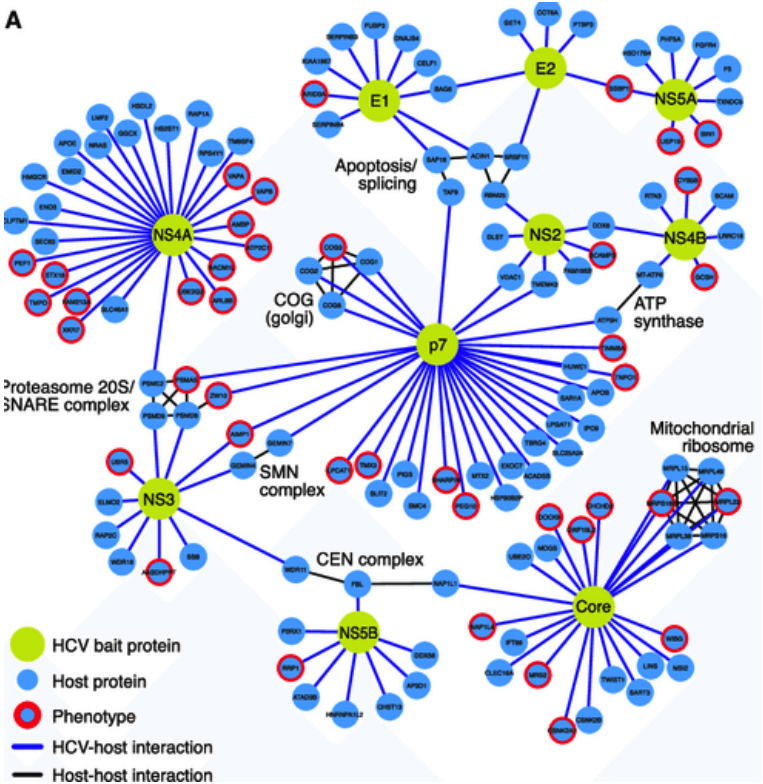
- Generate a concise summary based on connected terms

Host-viral protein interaction

- Two node types: human, HIV
- Two edge types
- Node attribute: affected by infection

- Propose mechanisms underlying the effect of infection

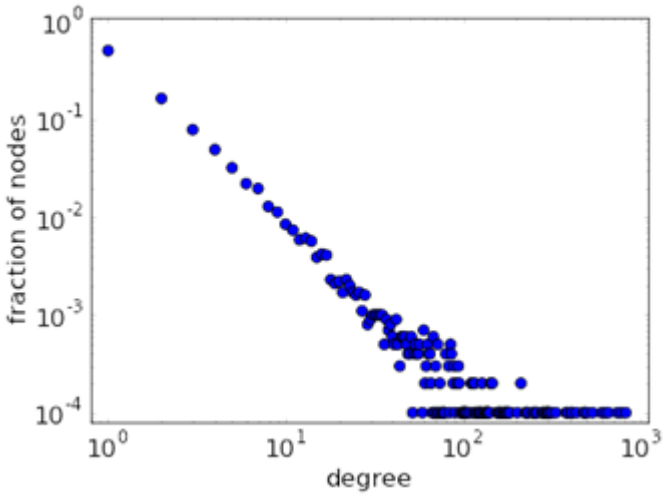
- Prioritize targets for antibody design



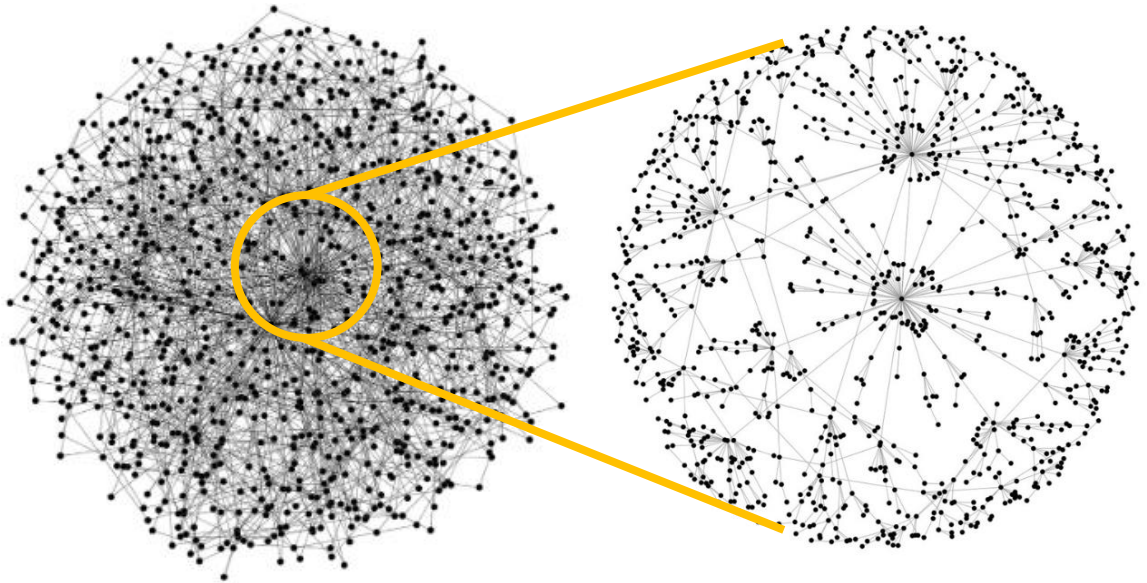
Source: Ramage et al. Mol Cell (2015)

Properties of real-world networks

Scale-free property



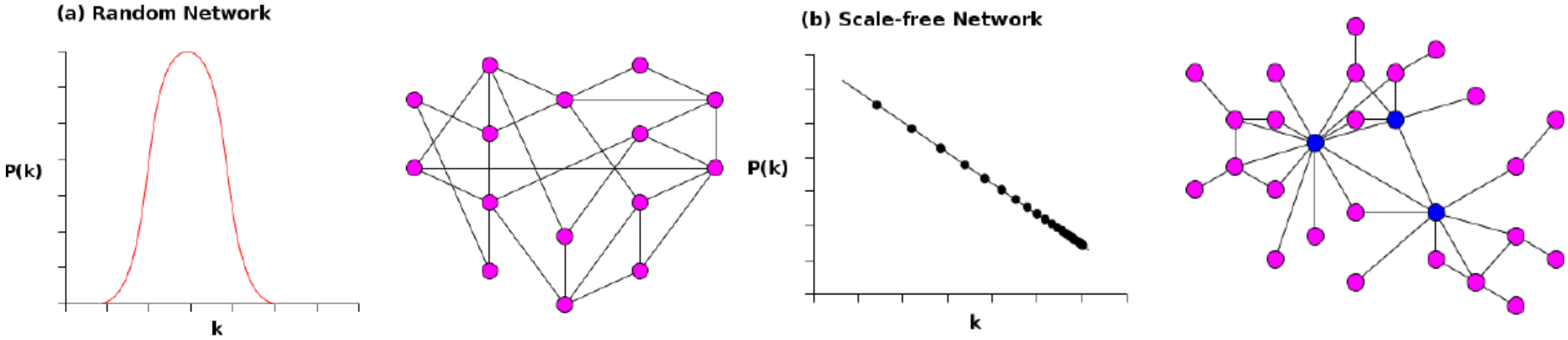
Source: mathinsight.org



Source: flickr.com & pacojariego.me

- Power law: P(node connected to k edges) ~ $1/k^{n}$
- Same local structure as global structure
  - Node-edge distribution

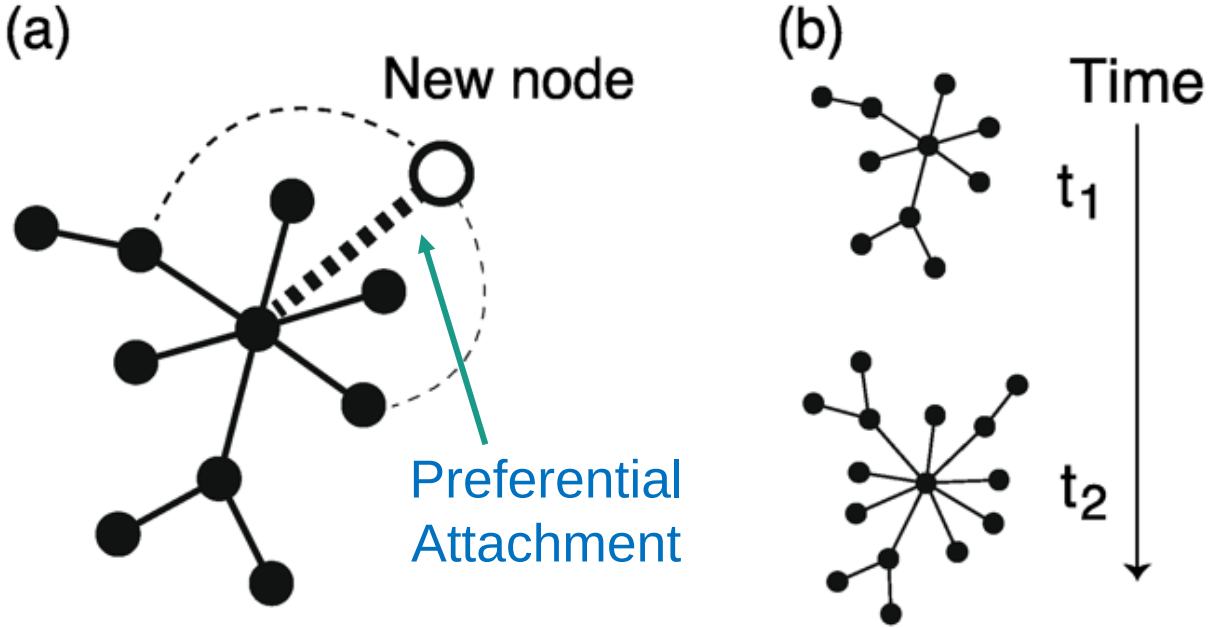
Hub and small-world propery



Source: Segura-Cebrera *et al*. Analysis of Protein Interaction Networks to Prioritize Drug Targets of Neglected-Diseases Pathogens

- Few nodes connected with many edges act as short-cut for traffic through the network
  - Transcription factors in biological networks
  - Social influencers on internet

Real-world networks from a simple mechanism

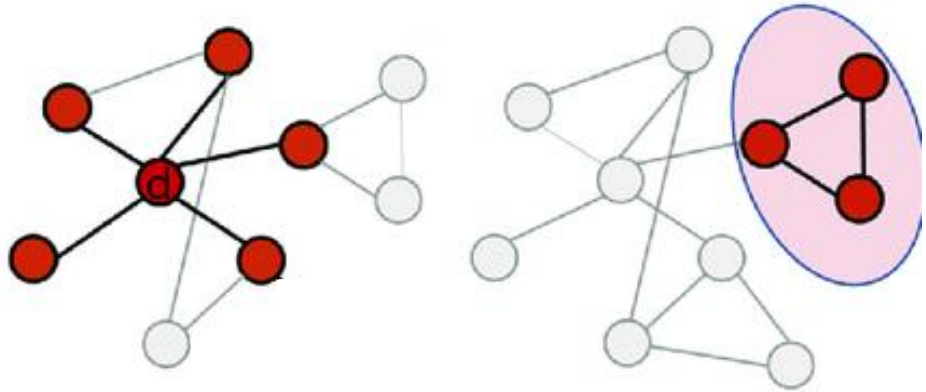


Takemoto. Metabolites, 2:429-457 (2012)

- New edges prefer to attach to existing nodes with already many edges

Topological properties

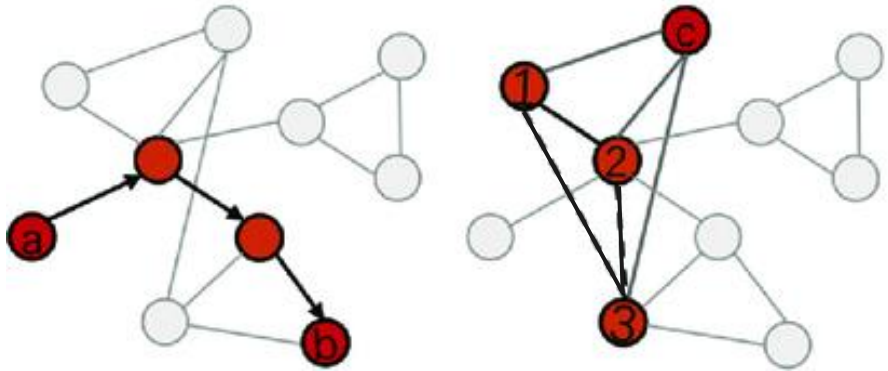
Degree and complete subgraph



Cai and Niu. Dev Cog Neuroscience (2018)

- Degree = number of edges connected to a certain node
- Clique, complete subgraph = region of a network whose nodes are fully connected with $\frac{n(n-1)}{2}$ edges

Path and clustering coefficient

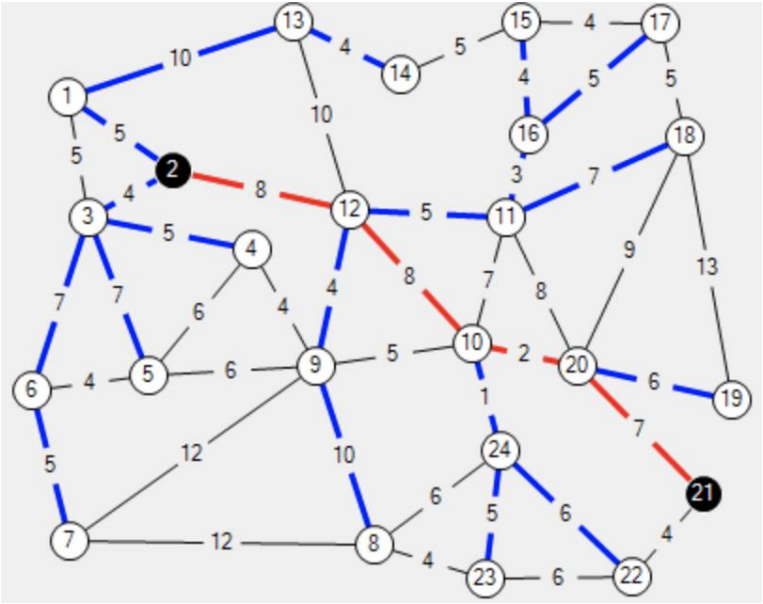


Cai and Niu. Dev Cog Neuroscience (2018)

- Path = connection from one node to another through several edges
  - Serve as distance between nodes on the network
- Clustering coefficient = proportion of neighbors that are also connected
  - Indicate the extent of local connectivity / redundancy of the network

Connectivity measures and interpretation

Importance of of path and path length



Source: Csharphelper.com

- Network operates by transmitting signals from one node to another
- Nodes frequently involved in the transmission are important
- **Model 1**: Shortest paths only
- **Model 2**: Weighted across all paths

Network as flows

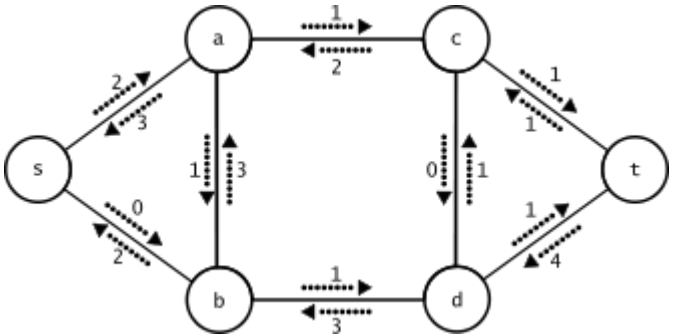


Image from https://en.wikipedia.org/wiki/Flow_network

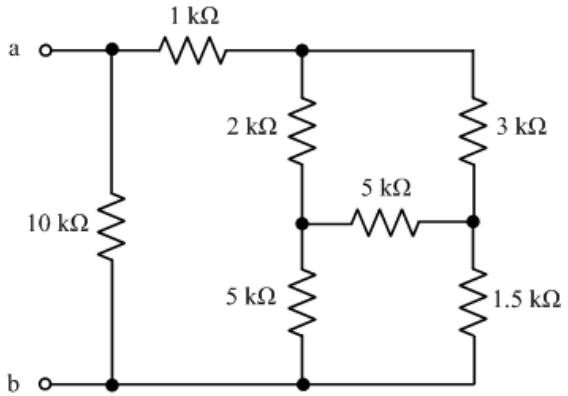
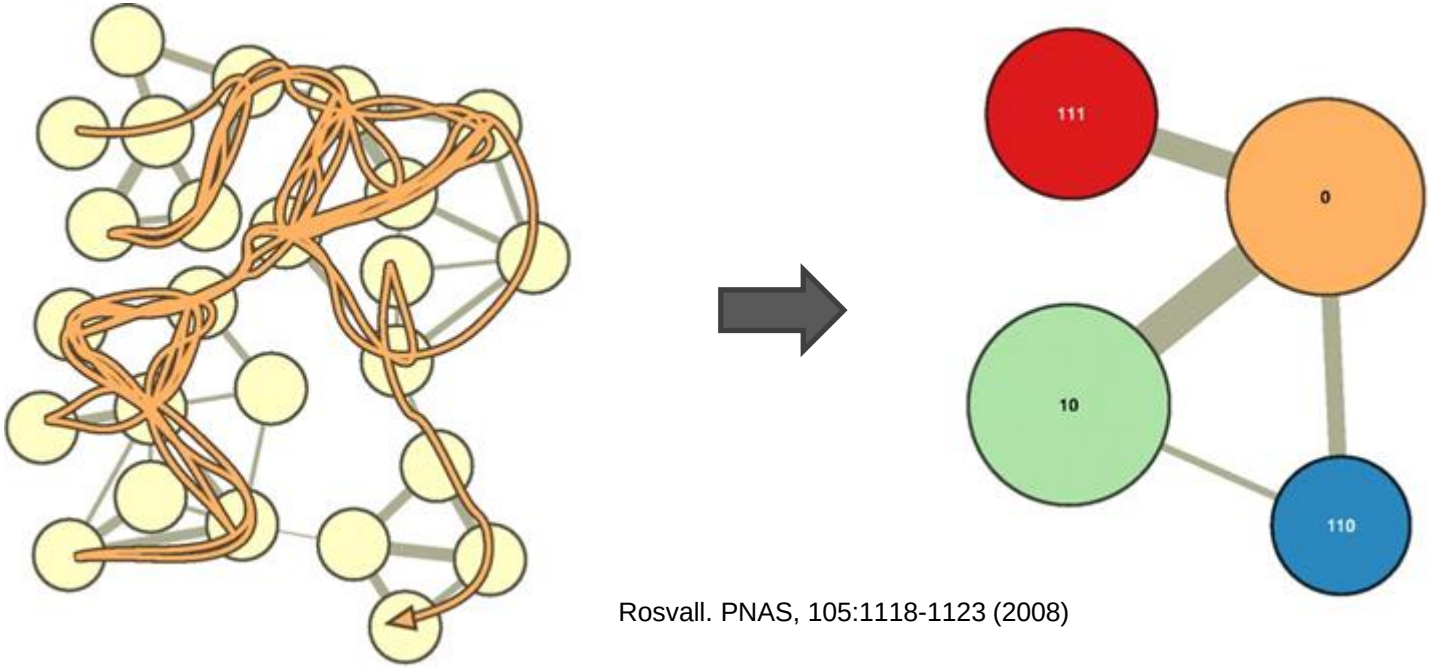


Image from http://www.rose-hulman.edu/CLEO/browse/?path=1/2/79/91/92/19

- Signal as fluid flowing through pipes with various diameters
- Signal as electric current flowing through circuit with various resistances
- Study the dynamics and stationary states with simulations

Network as random walks

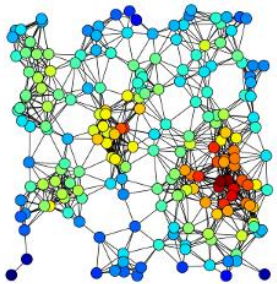


Rosvall. PNAS, 105:1118-1123 (2008)

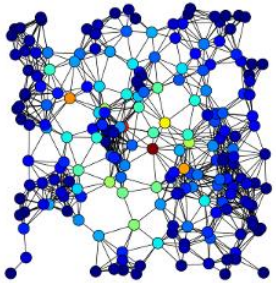
- Discrete particles travel from node to node with probability (edge weight)

Centrality scores

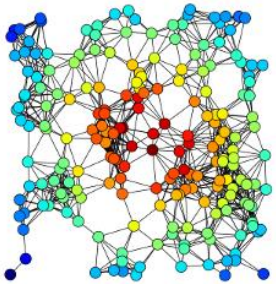
Degree



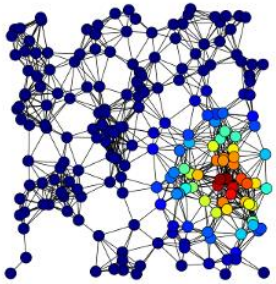
Betweenness



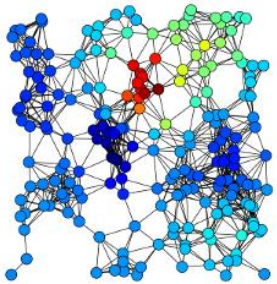
Closeness



Eigenvector



Katz



Harmonic

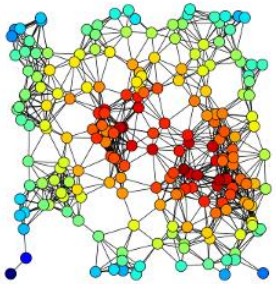
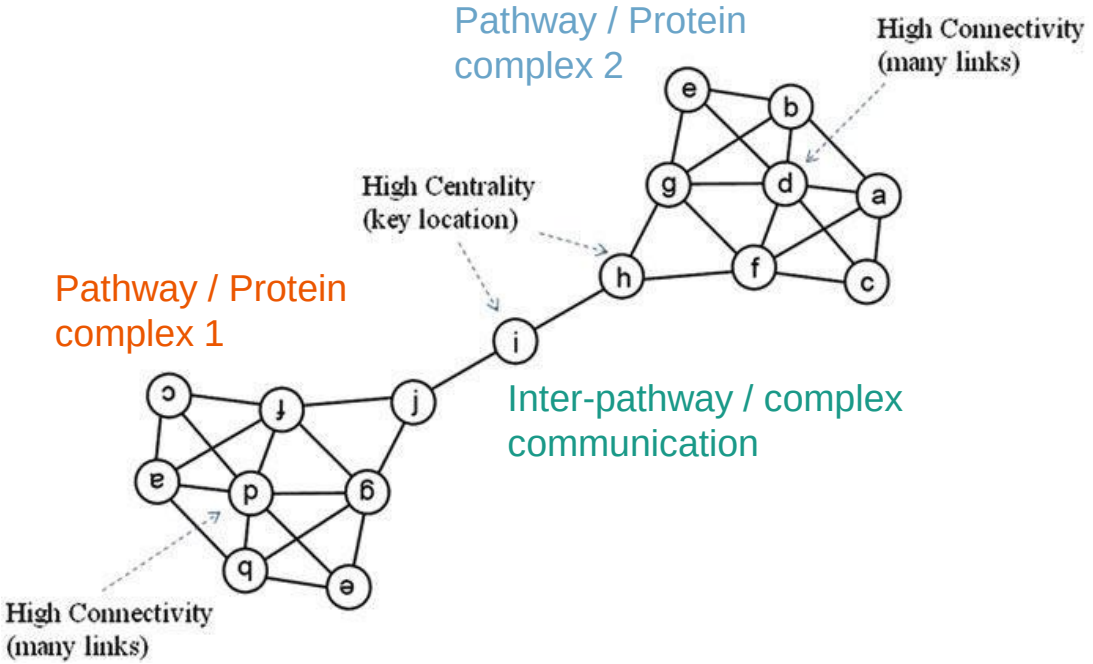


Image from https://en.wikipedia.org/wiki/Centrality

- Indicate the importance of a node in the context of the connectivity of the network
- Degree = local connectivity
- Betweenness = fraction of shortest paths
- Closeness = inverse distance to other nodes

Biological interpretation

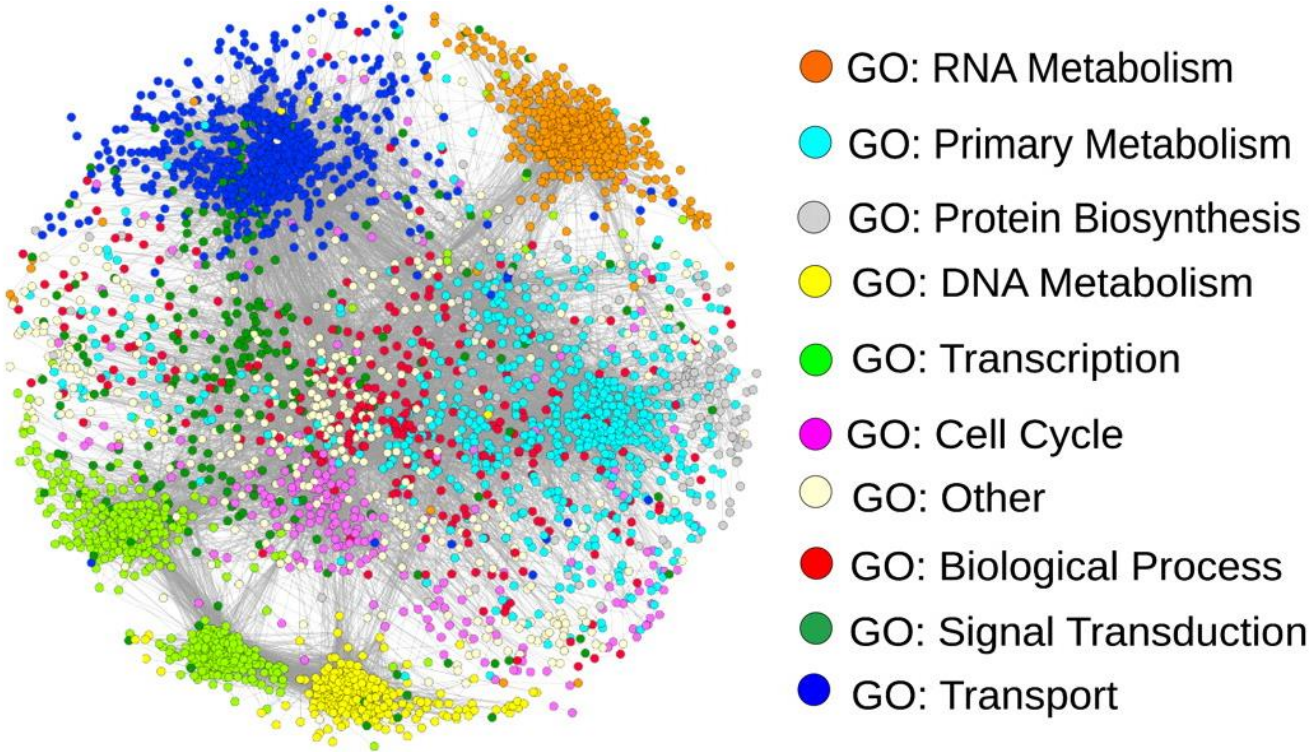


Natapov and Fisher-Gewirtzman. Smart City Networks (2017)

- **Low degree, High betweenness** = connect multiple functional pathways
- **High degree, Low betweenness** = core protein of a complex, transcription factor with multiple downstream targets

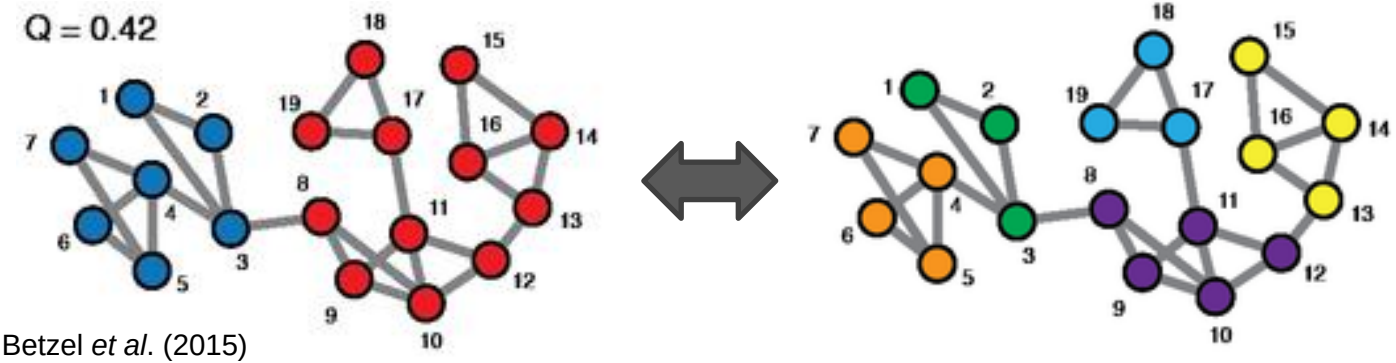
Network clustering

Dissect local characteristics

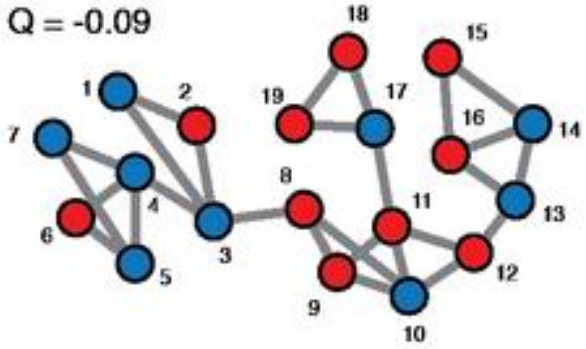


Pancaldi *et al*. G3: Genes, Genomics, Genetics, 2:453-497 (2012)

Modularity score



Betzel *et al*. (2015)



- Number of within-cluster edges compared to expectation (based on number of nodes and global number of edges)
- Multiple resolutions

A simple modularity score

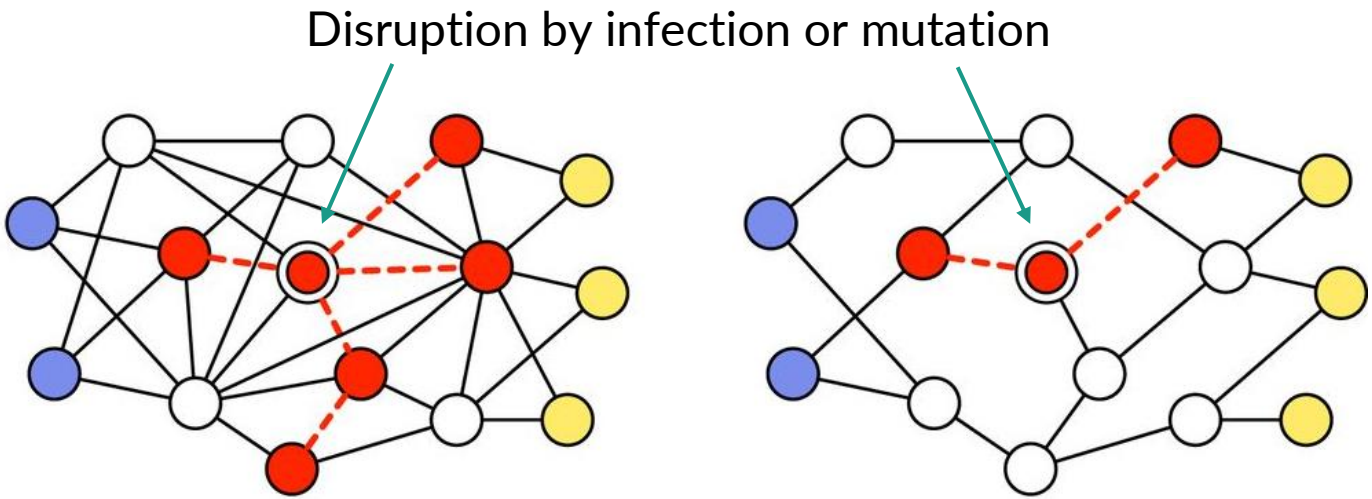
- Node N_i with degree d_i
- Node N_j with degree d_j

- Edge weights are all 1
- P(edge between N_i and N_j by chance) ~ d_i d_j / 2 x # edges

- Modularity score of a cluster of nodes (N_1, N_2, …, N_n) is then:
  - Q = # within-cluster edges $- \sum_{i,j} \frac{d_i d_j}{2e}$

Applications of network analyses

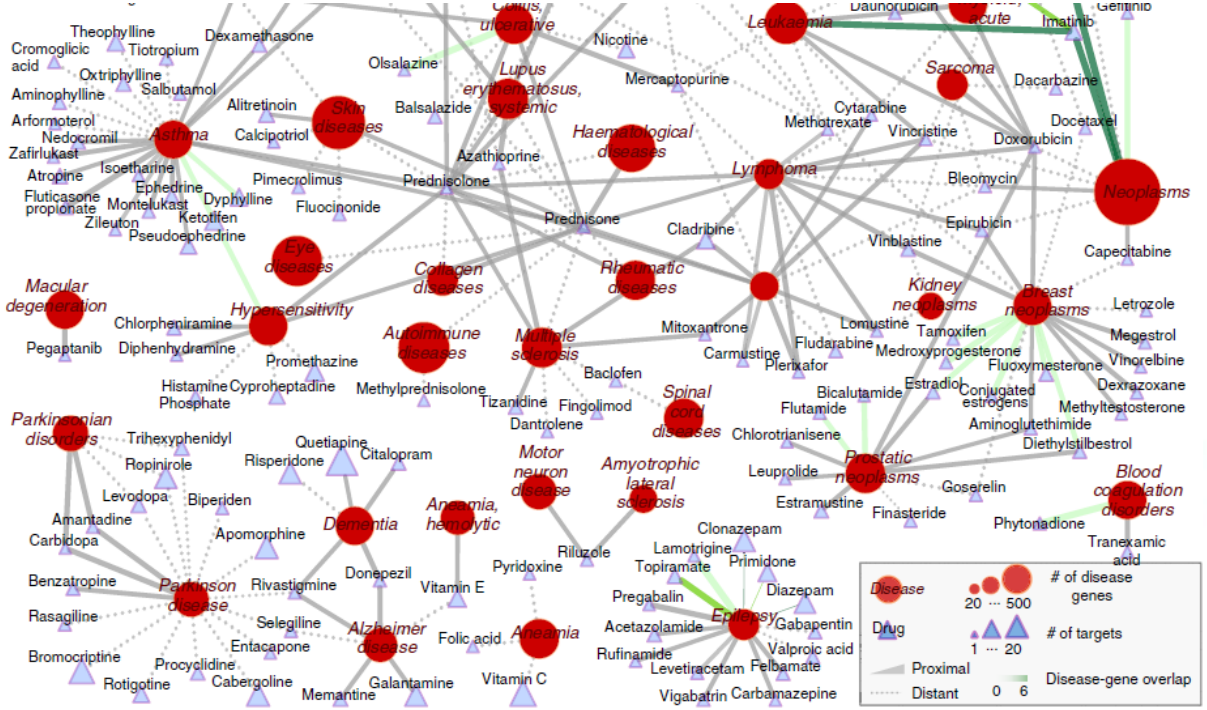
What if nodes/edges were removed?

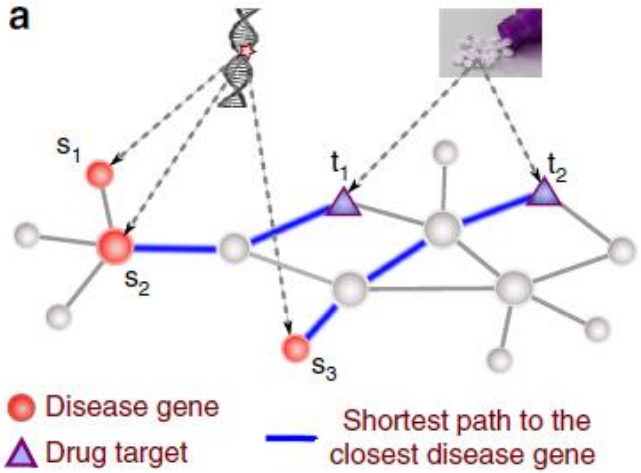


Navlakha *et al*. J of the Royal Society Interface, 11 (2014)

- Analysis of network-level changes induced by node/edge-level changes
- Complement centrality scores

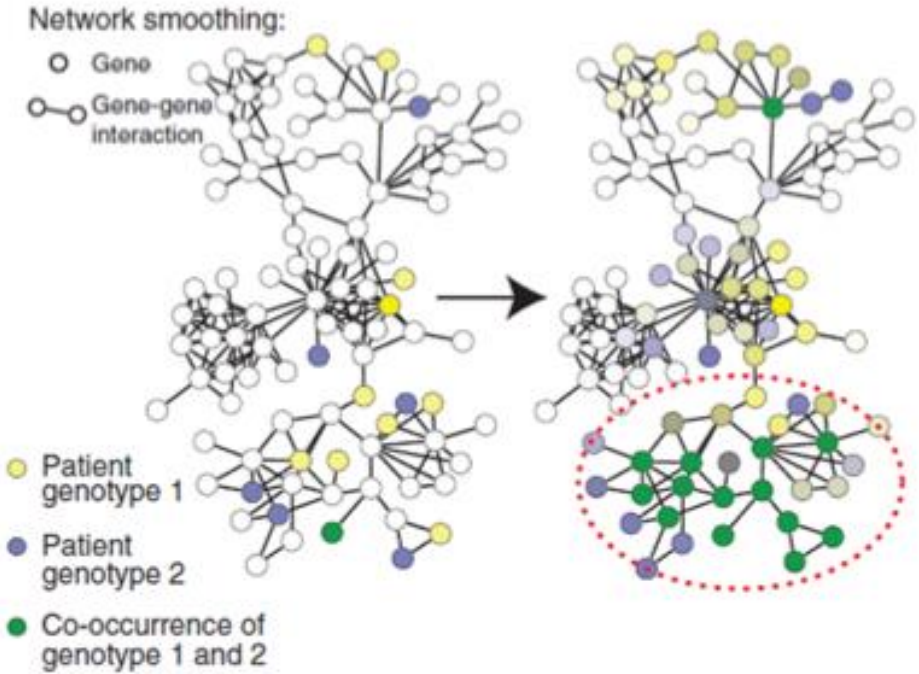
Linking drug to disease via gene network





Guney *et al*. Nature Comm, 7:10331 (2016)

Network-based patient stratification



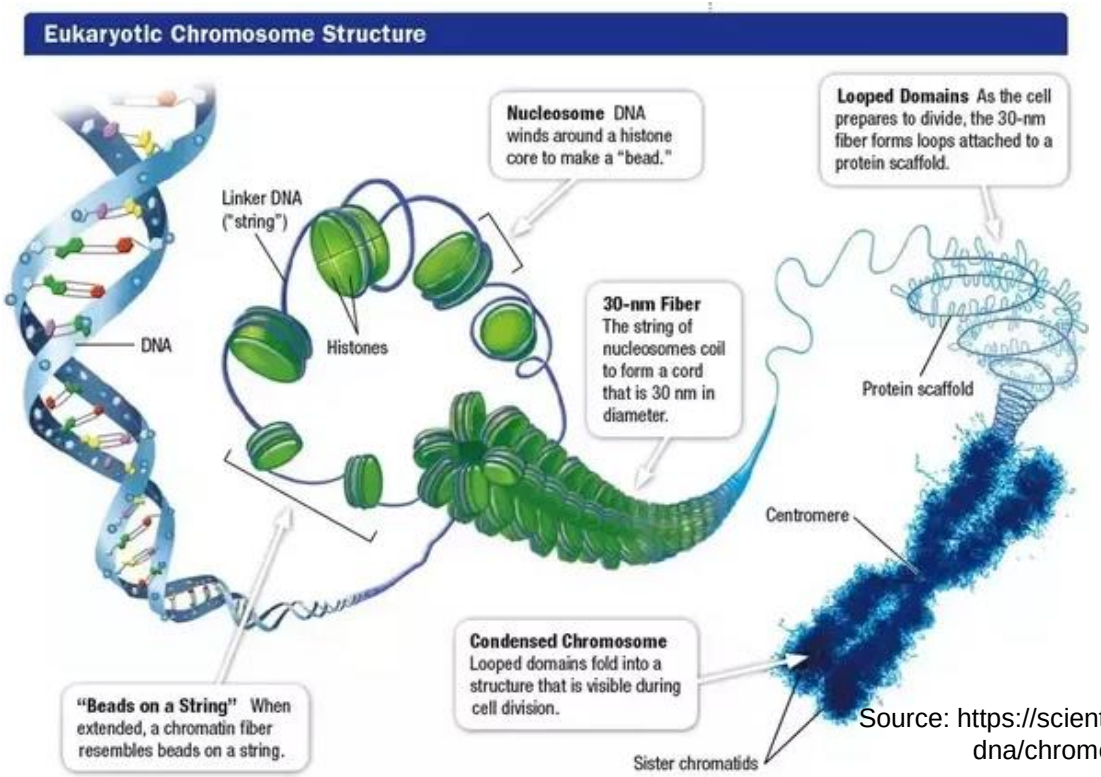
- Different patients have different mutation profiles
- Different mutation profiles may have similar impacts on gene-gene network
- Identify commonly affected gene subnetworks

Hofree *et al*. Nature Methods, 10:1108-1115 (2013)

Part 2: Chromatin organization

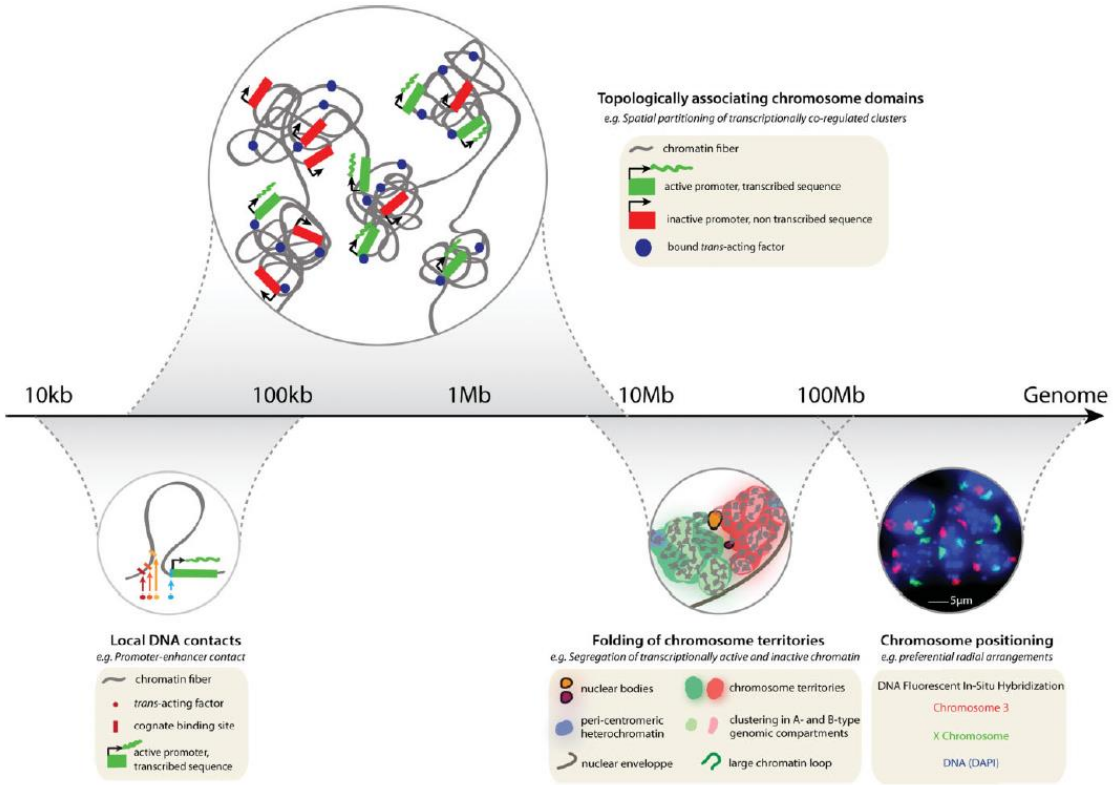
- Many mechanisms of gene regulation involve chromatin (epigenetics)
- 3D folding of chromatin is quite well organized and conserved across evolution
- A/B compartment and topologically associating domain (TAD)

Chromatin folding for packaging



Source: https://scientificallytalking.com/what-is-dna/chromosome-structure/

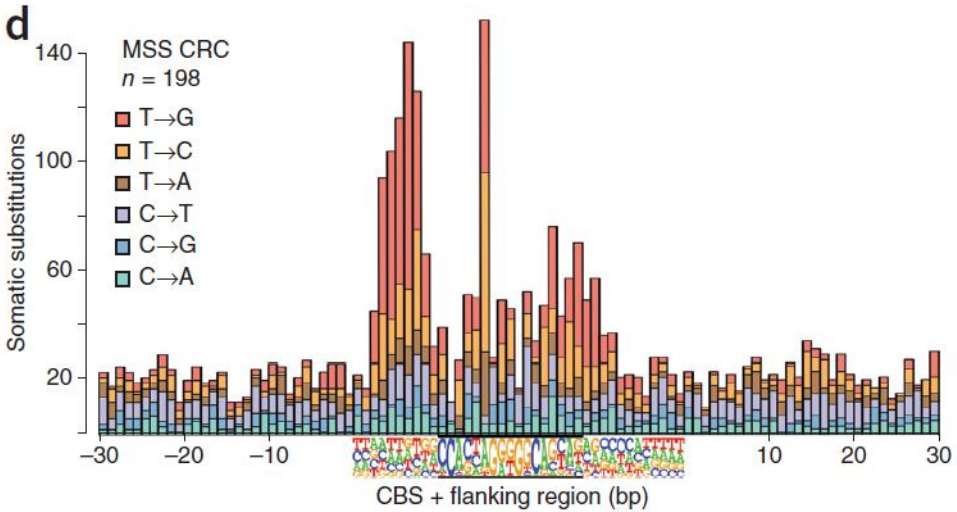
Hierarchical organization of chromatin

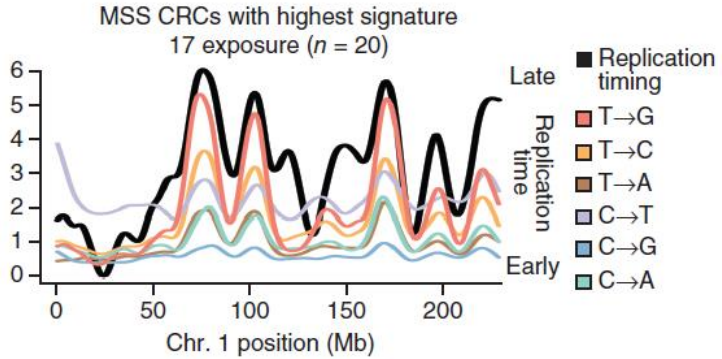


- Organized chromatin is easy to split during cell division
- Topologically associating domain (TAD) = local chromatin folds
- Enhancer looping

Source: Nora *et al*. Bioessay (2013)

Cancer subtype with destabilized chromatin

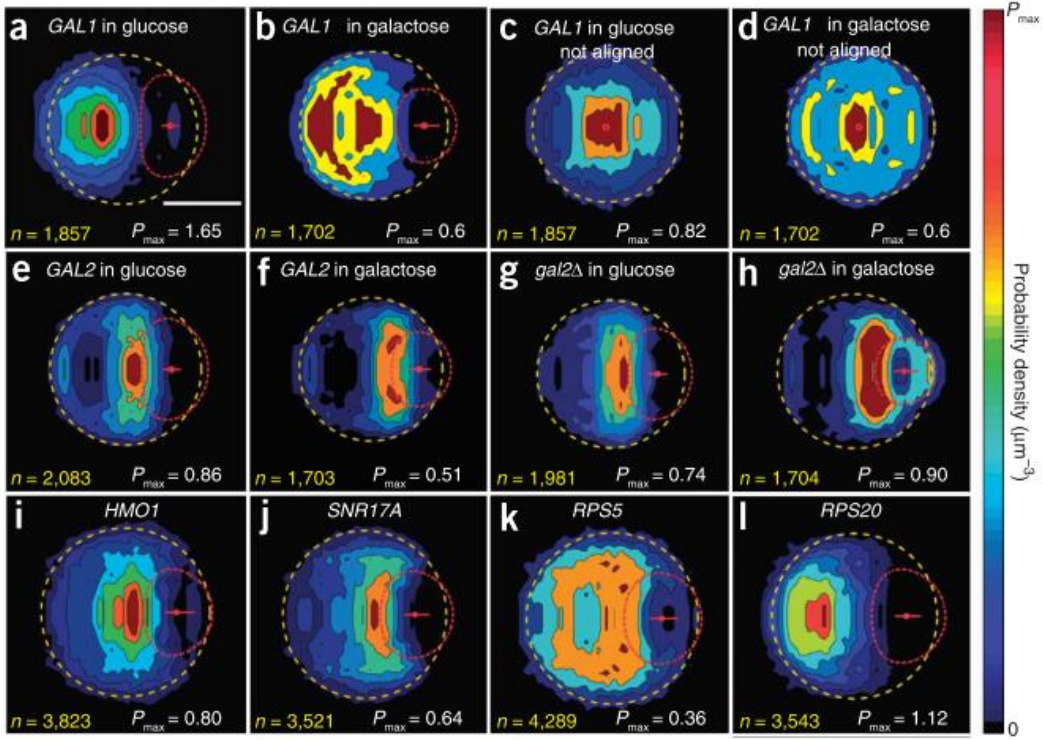


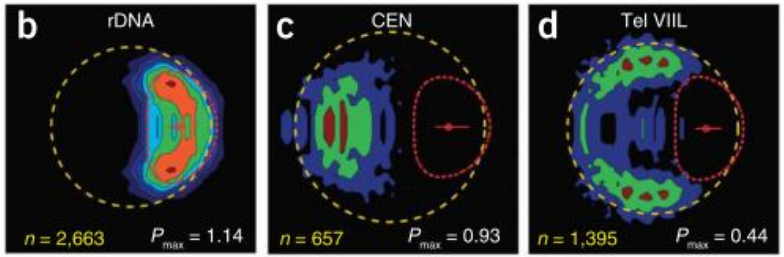


Katainen *et al*. Nature Genetics (2015)

- CBS = CTCF binding site on chromosome to regulate the folding
- Mutations on CBS disrupt chromatin folding
  - Widespread dis-regulations of gene expression

Chromain and gene territories

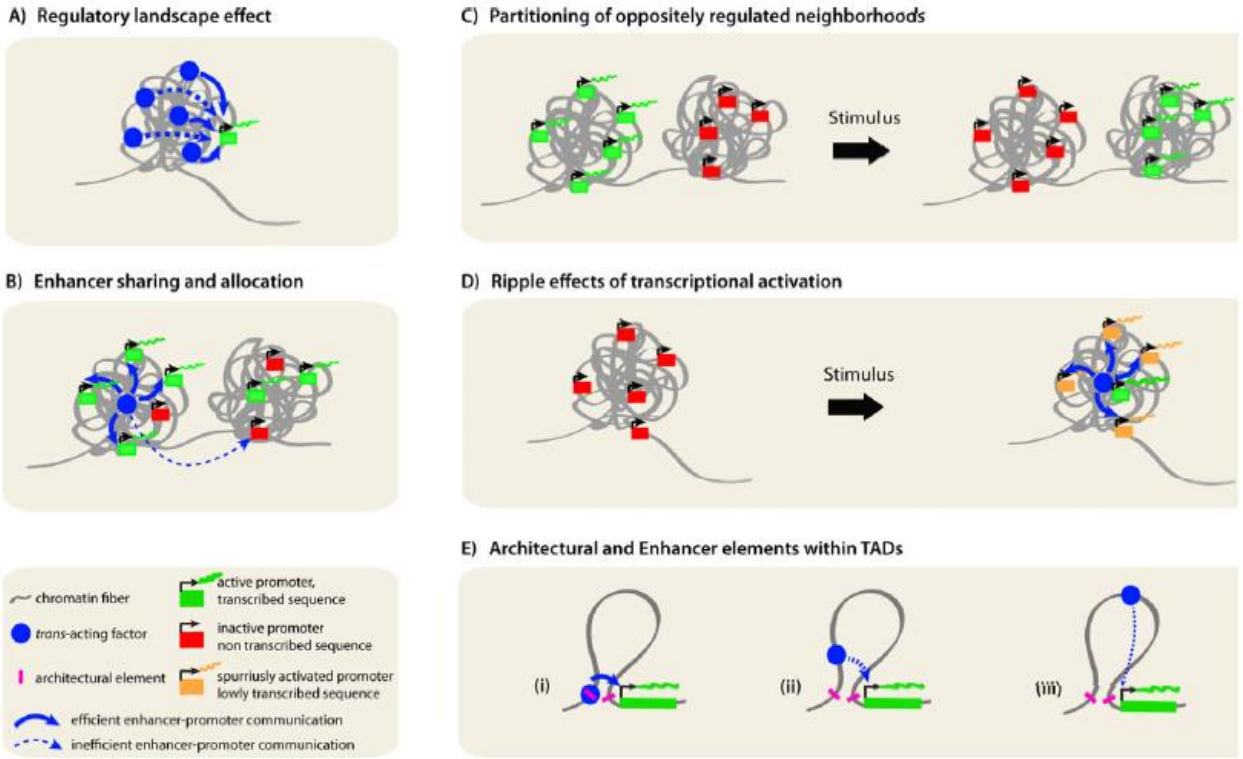




- Chromatin features, such as rDNA, centromere, and telomere, and some genes occupy specific regions in the nucleus

Source: Berger *et al*. Nature Methods (2008)

Biological implications of gene territories

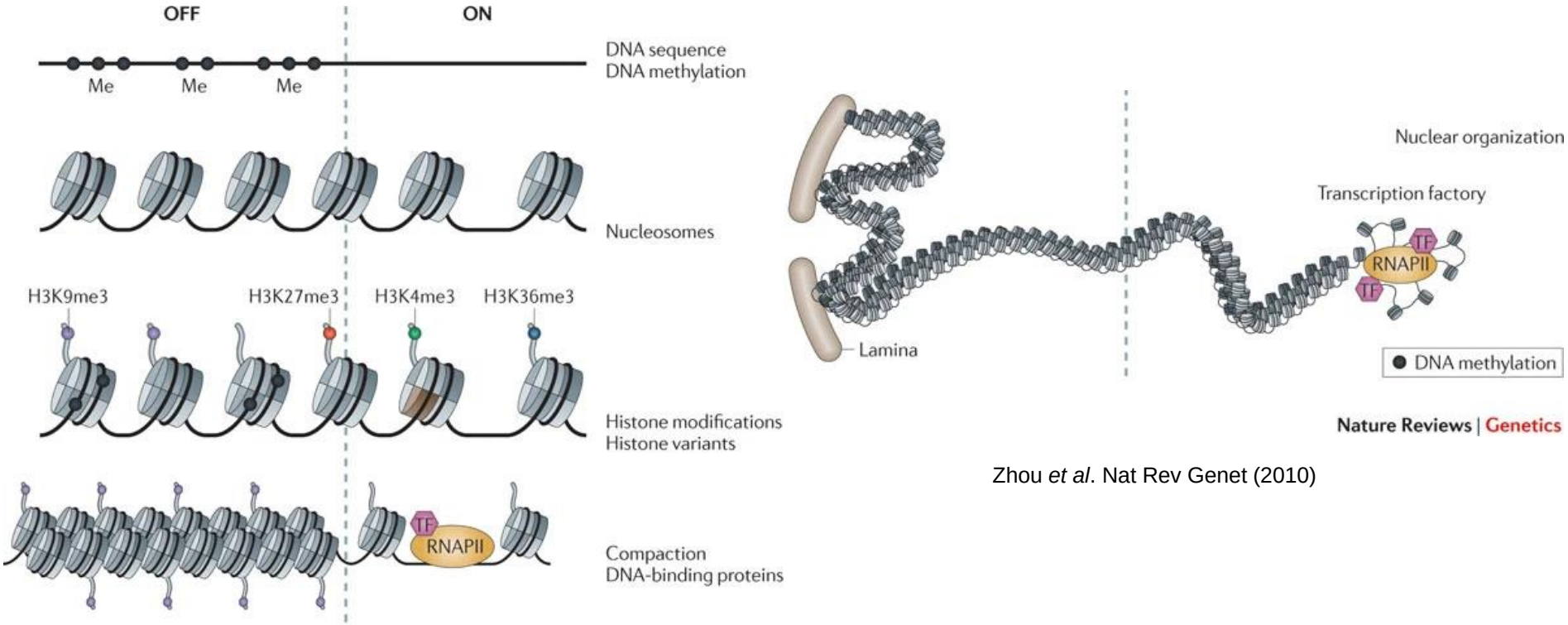


- Localization of co-regulated genes
- Sharing of transcription factors and enhancer

Nora *et al*. Bioessay (2013)

Chromatin assays

Chromatin and epigenomics assays



Zhou *et al*. Nat Rev Genet (2010)

Bisulfite sequencing analysis

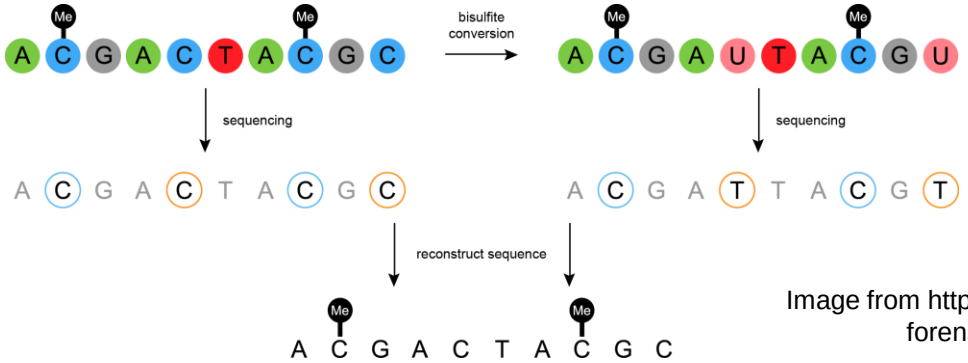
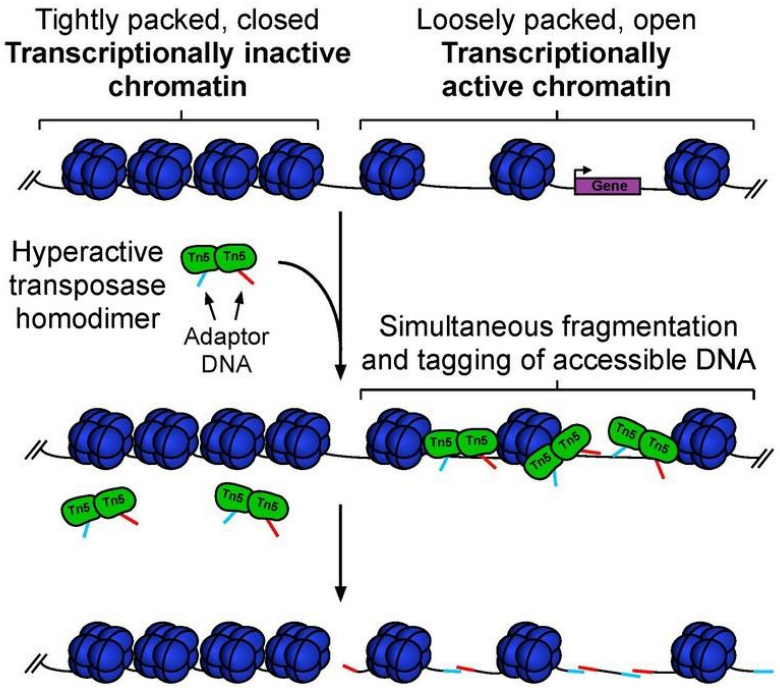
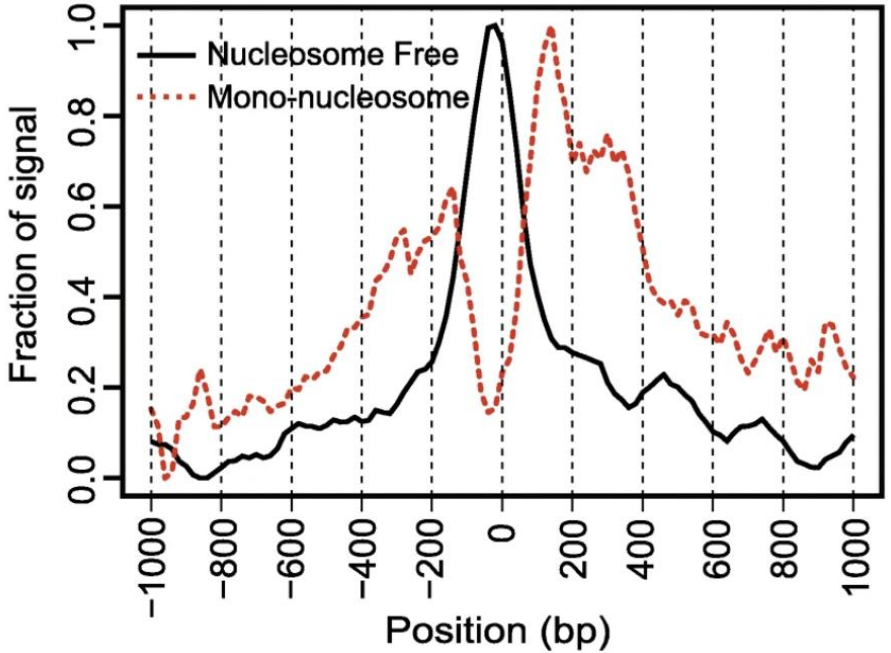


Image from http://www.atdbio.com/content/20/Sequencing-forensic-analysis-and-genetic-analysis

- Strategy 1: Convert C in reference genome to T
  - Use regular aligner + compare match to C and T versions
- Strategy 2: Adjust mismatch score for C-T
  - P(random C-T mismatch) scales with P(T in read)
  - P(C-T bisulfite) scales with P(non-methylated)

ATAC-seq





Yan *et al*. Genome Biology (2020)

ChIP-seq and peak calling

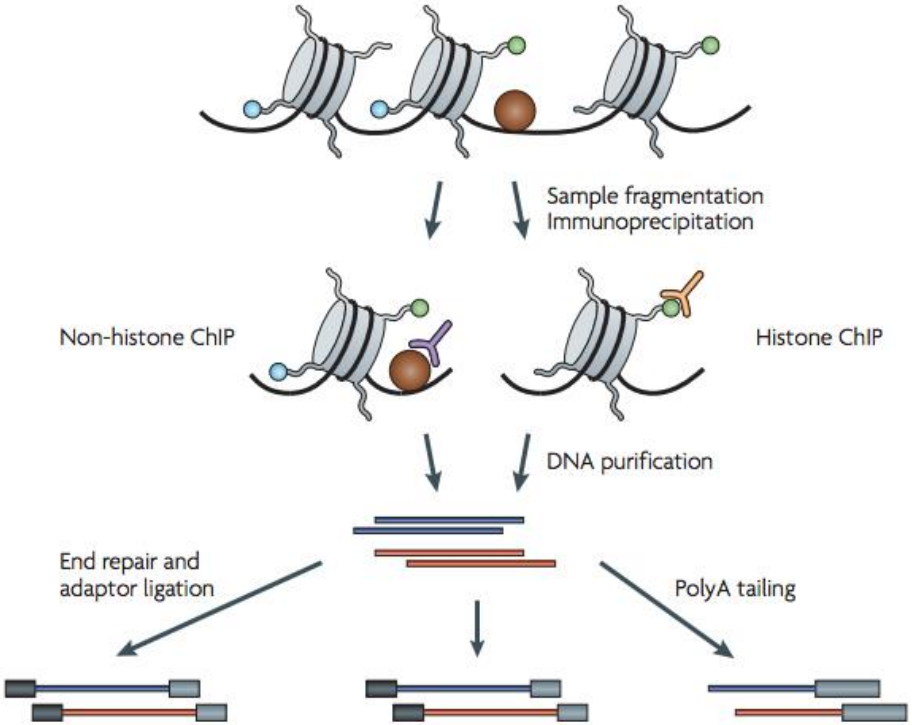
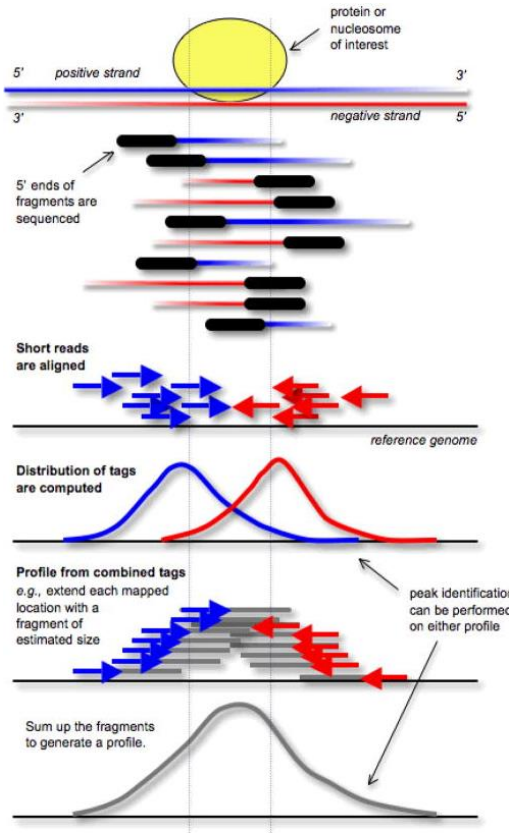


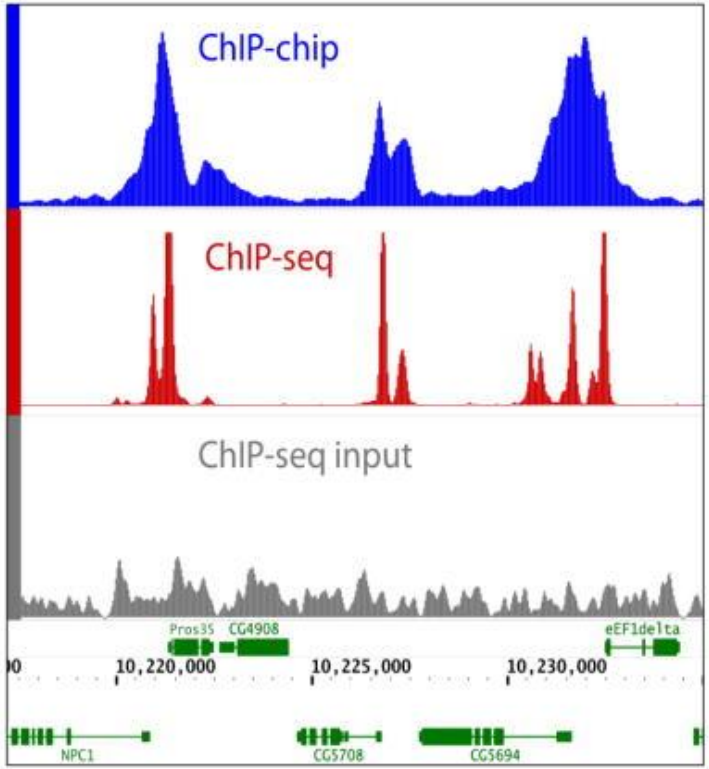
Image from http://informatics.fas.harvard.edu/chip-seq-workshop.html

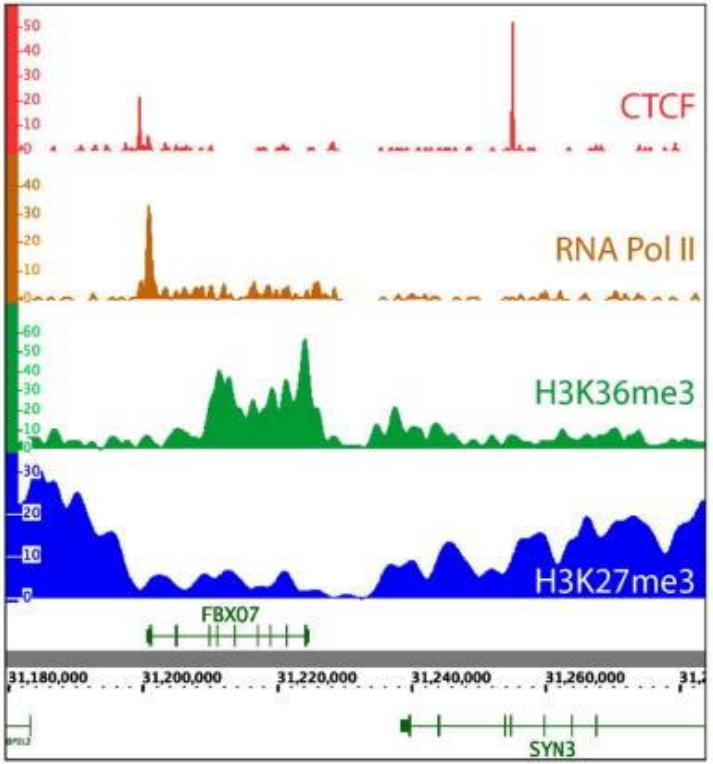


Park *et al.* Nat Rev Genet 10:669-680 (2009)

Broad and sharp peaks

Park *et al.* Nat Rev Genet 10:669-680 (2009)





ChIP-seq analysis

- Starts off like any sequencing dataset

- Peak calling and annotation
  - Linking peaks to genes
    - Nearest gene?
    - All genes within 2000 bp?
  - Functional enrichment

- Motif search
  - Find common DNA patterns across peaks
  - Possible TF binding sites

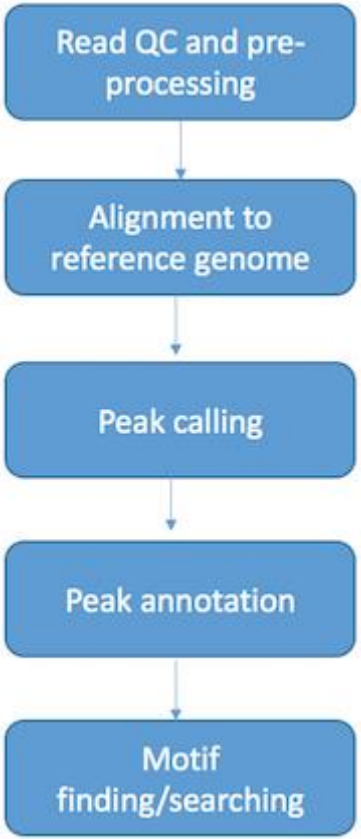


Image from http://informatics.fas.harvard.edu/chip-seq-workshop.html

Poisson model for peak calling

- Number of reads at each genomic locus ~ Poisson(λ)
  - λ varies and represent the background number of reads
  - Estimate λ from control samples (no immunoprecipitation)

- P-value for observing n reads at a locus with λ reads in the control
  - $\sum_{k=n}^{N} \frac{\lambda^k e^{-\lambda}}{k!}$

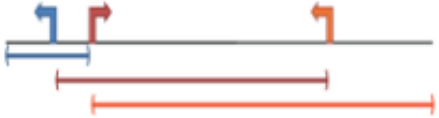
Peak annotation

(•) Basal plus extension

Proximal: 5.0 kb upstream, 1.0 kb downstream, plus Distal: up to 1000.0 kb

Gene regulatory domain definition: Each gene is assigned a basal regulatory domain of a minimum distance upstream and downstream of the TSS (regardless of other nearby genes). The gene regulatory domain is extended in both directions to the nearest gene's basal domain but no more than the maximum extension in one direction.

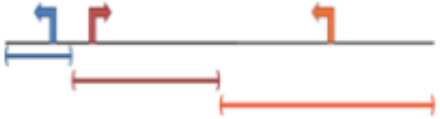
( ) Two nearest genes



within 1000.0 kb

Gene regulatory domain definition: Each gene is assigned a regulatory domain that extends in both directions to the nearest gene's TSS but no more than the maximum extension in one direction.

( ) Single nearest gene

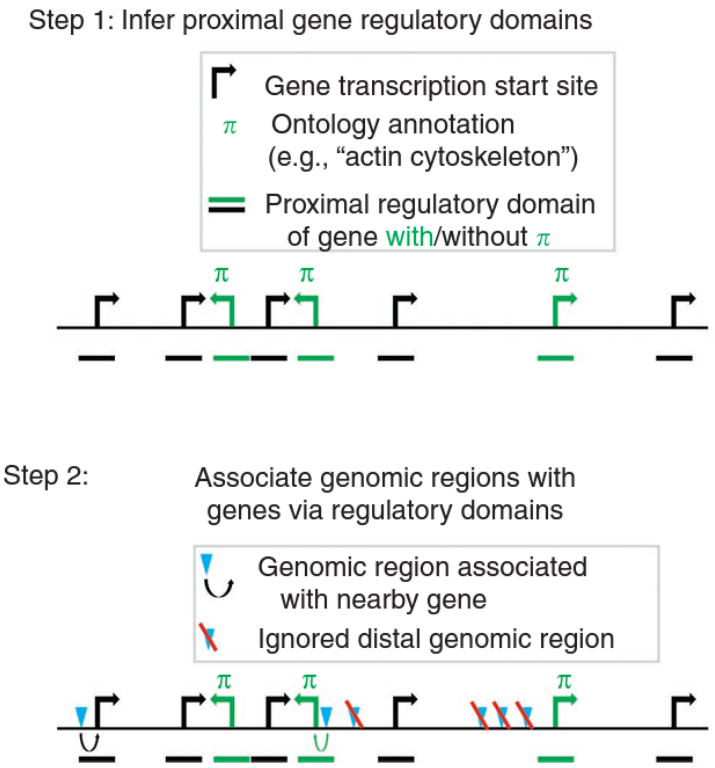


within 1000.0 kb

Gene regulatory domain definition: Each gene is assigned a regulatory domain that extends in both directions to the midpoint between the gene's TSS and the nearest gene's TSS but no more than the maximum extension in one direction.

Gene Transcription Start Site (TSS)

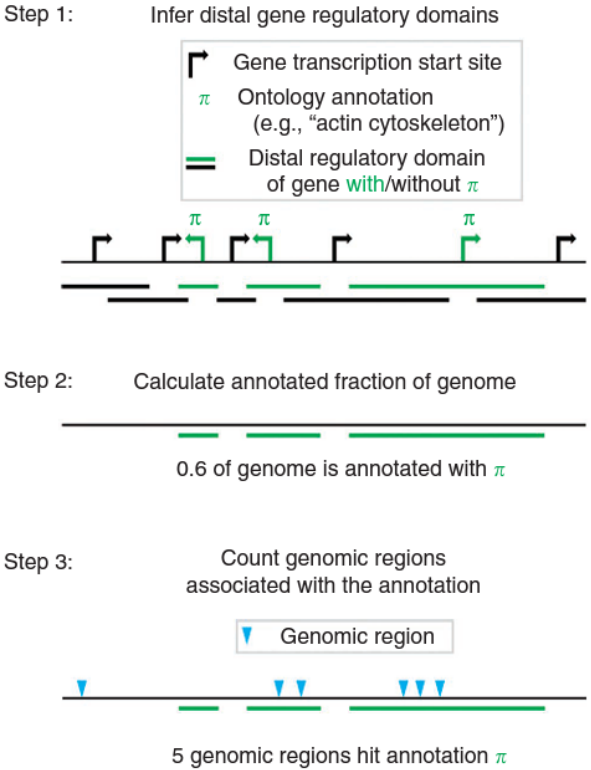
Overrepresentation analysis



- Are peaks located near genes with certain functionality?
- Use peak locations instead of differential expression
- Hypergeometric distribution

McLean *et al*. Nat Biotech, 28(5):495-501 (2010)

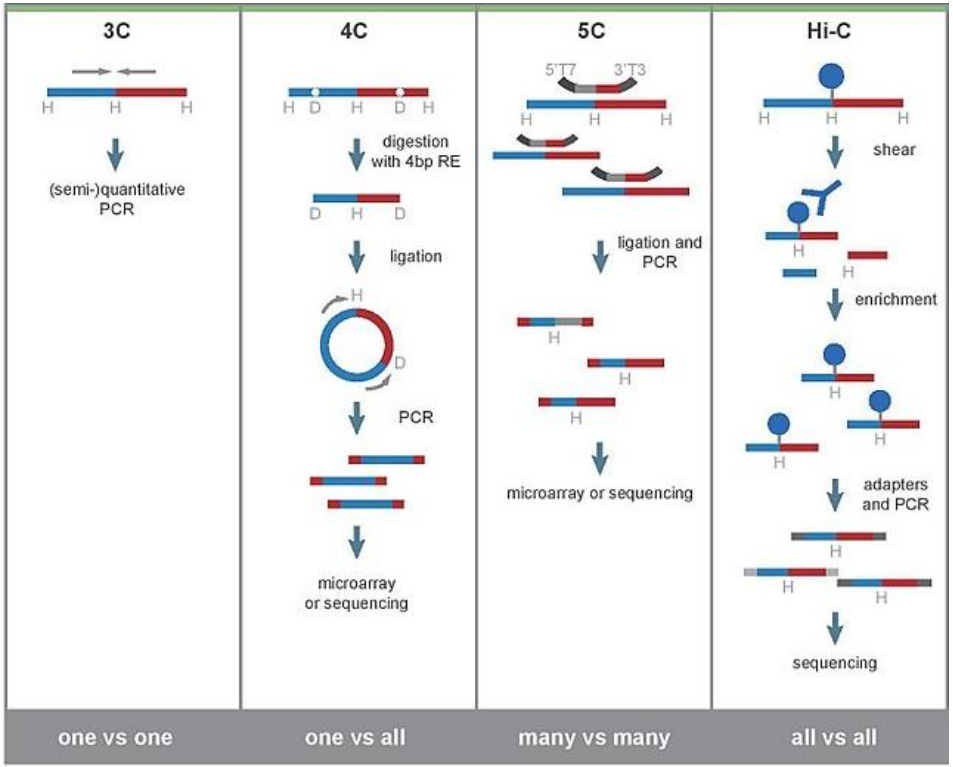
Overrepresentation analysis with binomial model

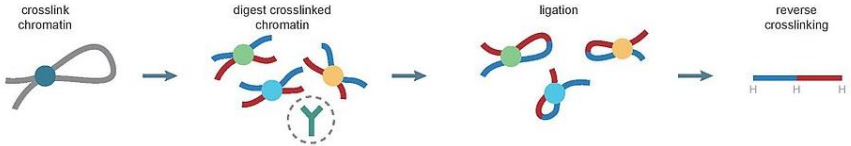


- Assign function to genomic regions surrounding the genes
  - Fraction of genome = success rate
- Identify number of peaks located in the annotated regions
  - Number of success
- Binomial distribution

Long-range chromatin interaction

Chromatin conformation capture





- Induce cross-link between proximal chromatins
- Generate hybrid DNA fragments
- Identify by high-throughput sequencing

Chromatin structure resolution

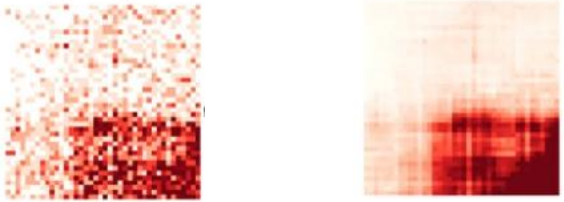
Article | Open Access | Published: 21 February 2018

Enhancing Hi-C data resolution with deep convolutional neural network HiCPlus

Yan Zhang, Lin An, Jie Xu, Bo Zhang, W. Jim Zheng, Ming Hu, Jijun Tang ✉ & Feng Yue ✉

Nature Communications **9**, Article number: 750 (2018) | Cite this article

Low-resolution High-resolution

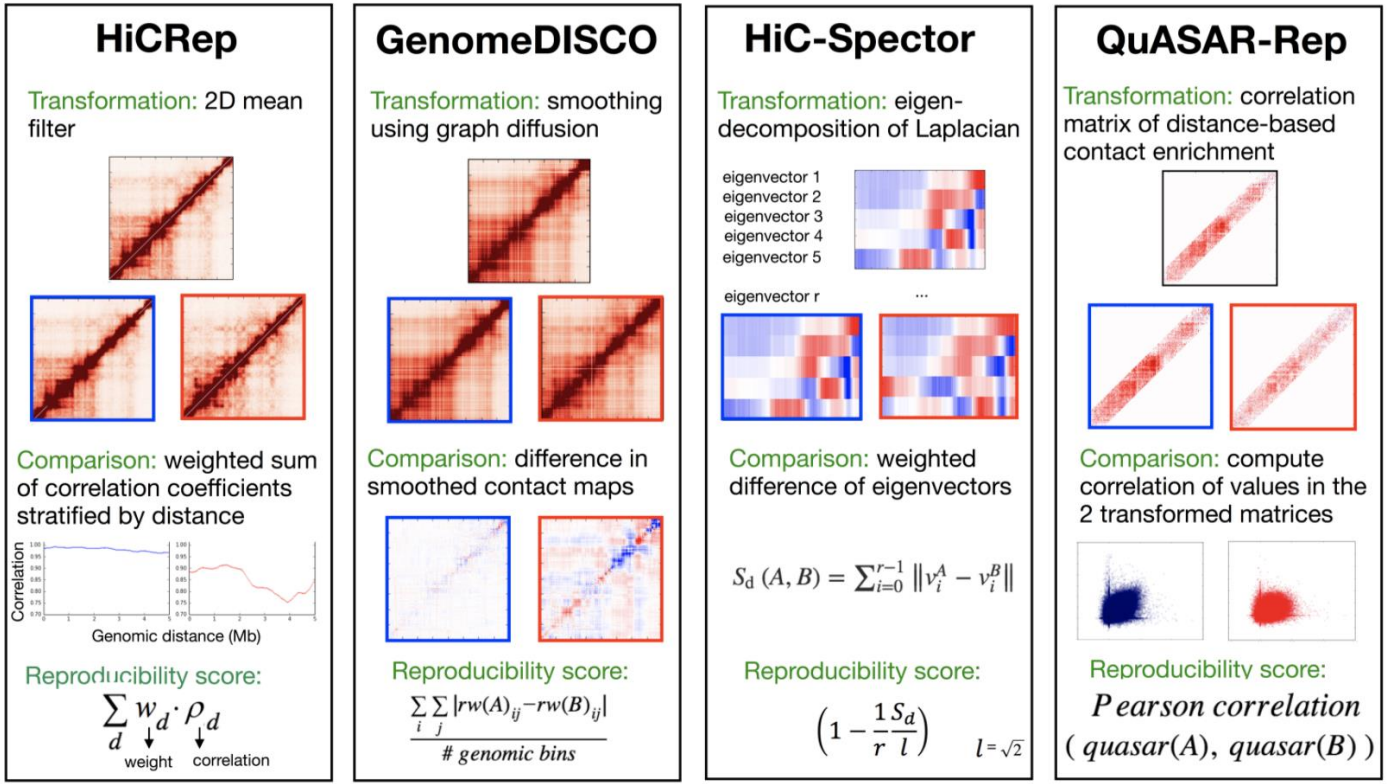


Zhang *et al.*, Nat Comm (2018)

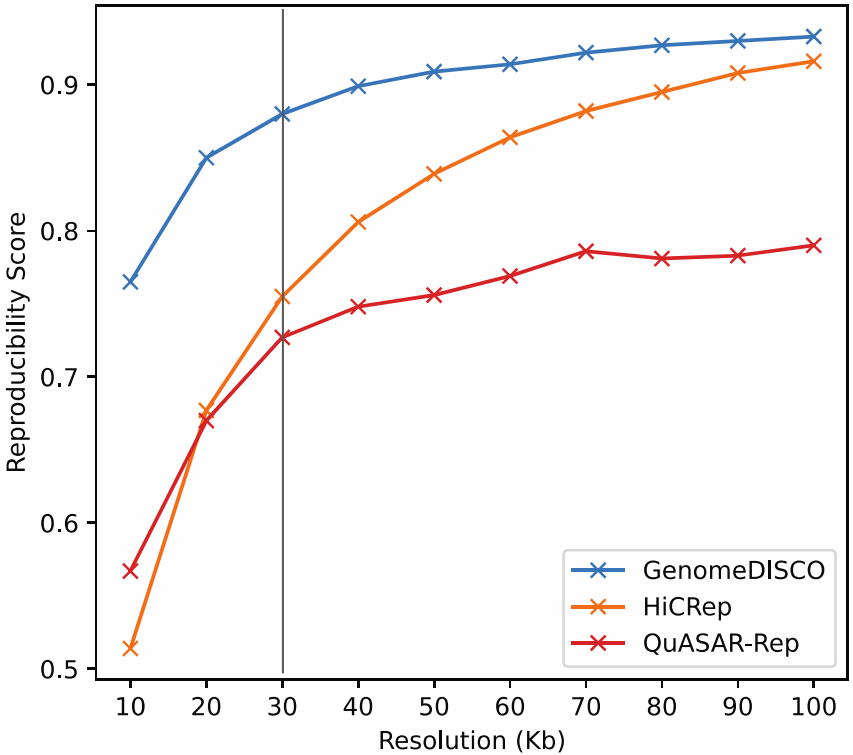
- Studying chromatin structure with Hi-C = filling an *N* x *N* matrix with read counts
- Requite N^2 read depths
- Consolidate genomes into bins
  - 10kb, 20kb, ..., 100kb
  - Larger bins = require less reads

QC for Hi-C data

Source: github.com/kundajelab/3DChromatin_ReplicateQC



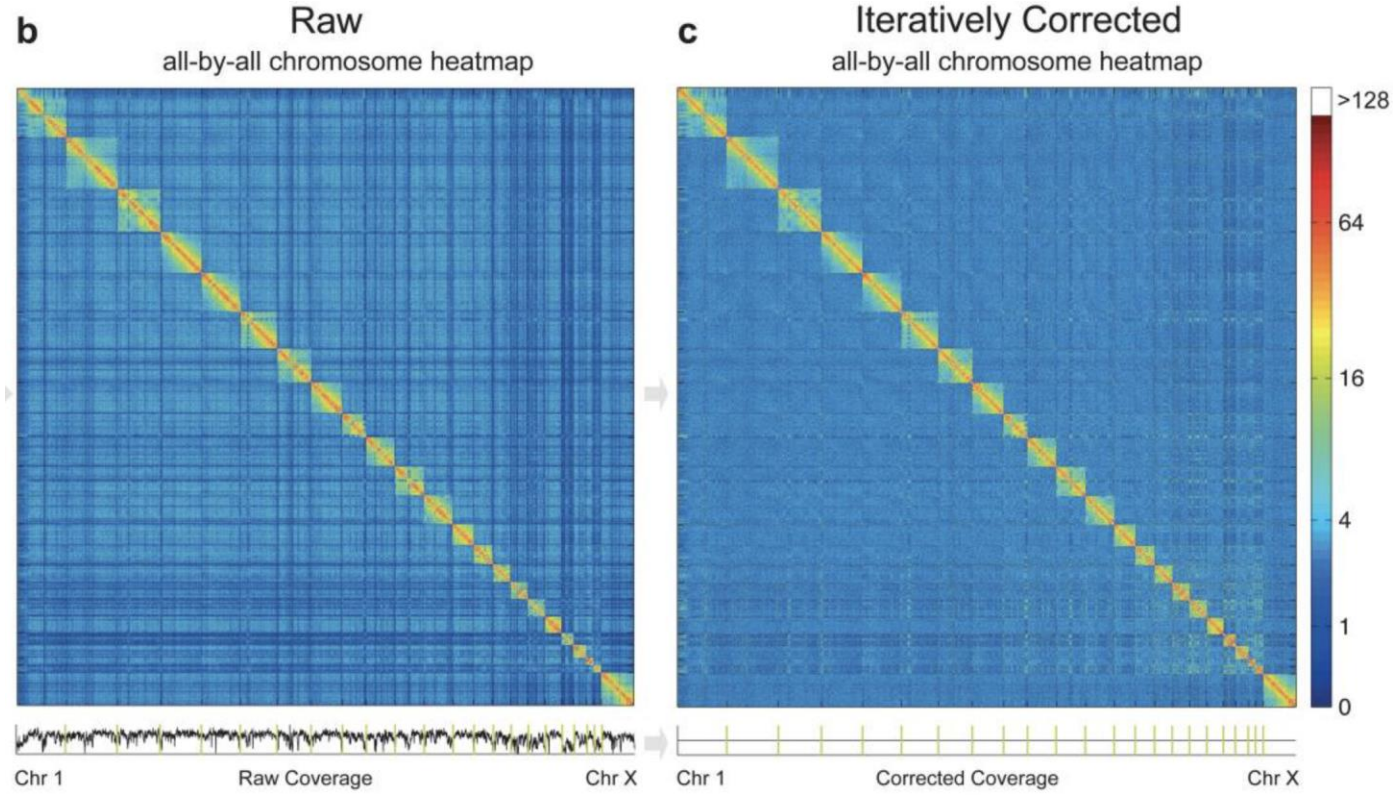
Finding appropriate resolution



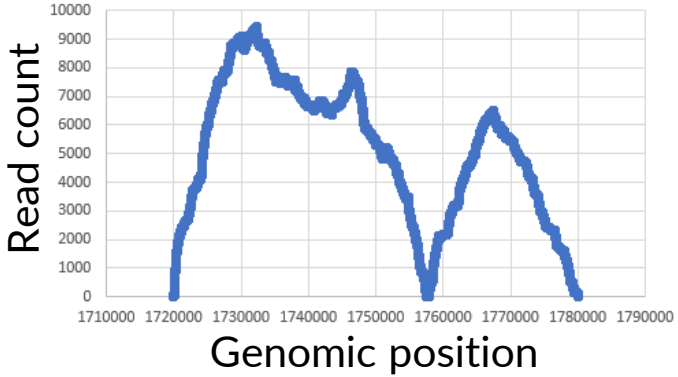
- Low read depth at high resolution → high variance across bootstraps
- Reducing resolution beyond a certain point → no benefit
- Typically split data in half

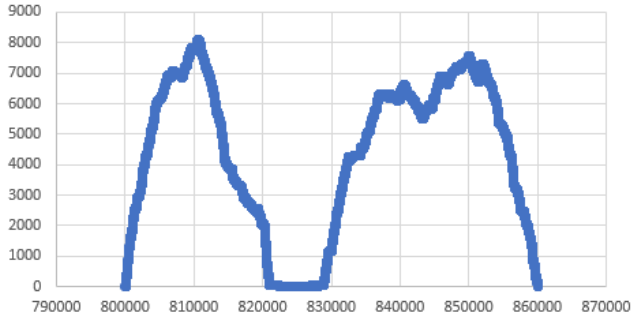
Iterative correction and eigenvector decomposition (ICED)

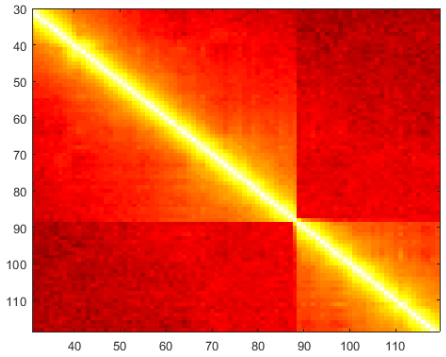
Imakaev et al. Nature Methods (2012)

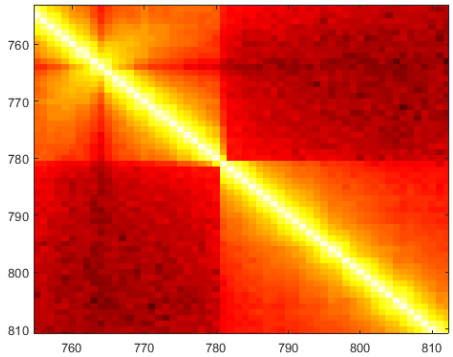


Hi-C identifies misassembled genomic region



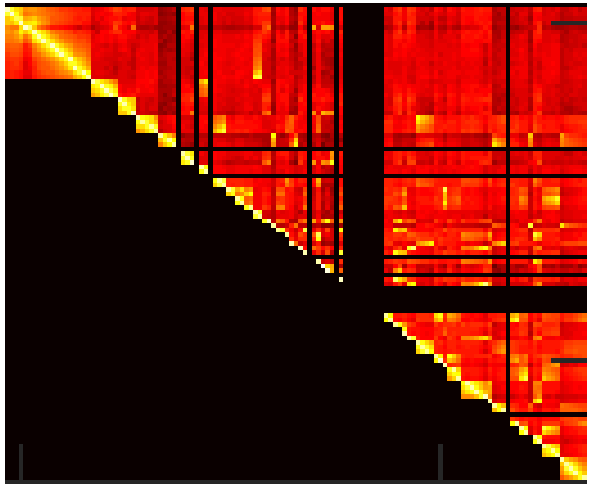


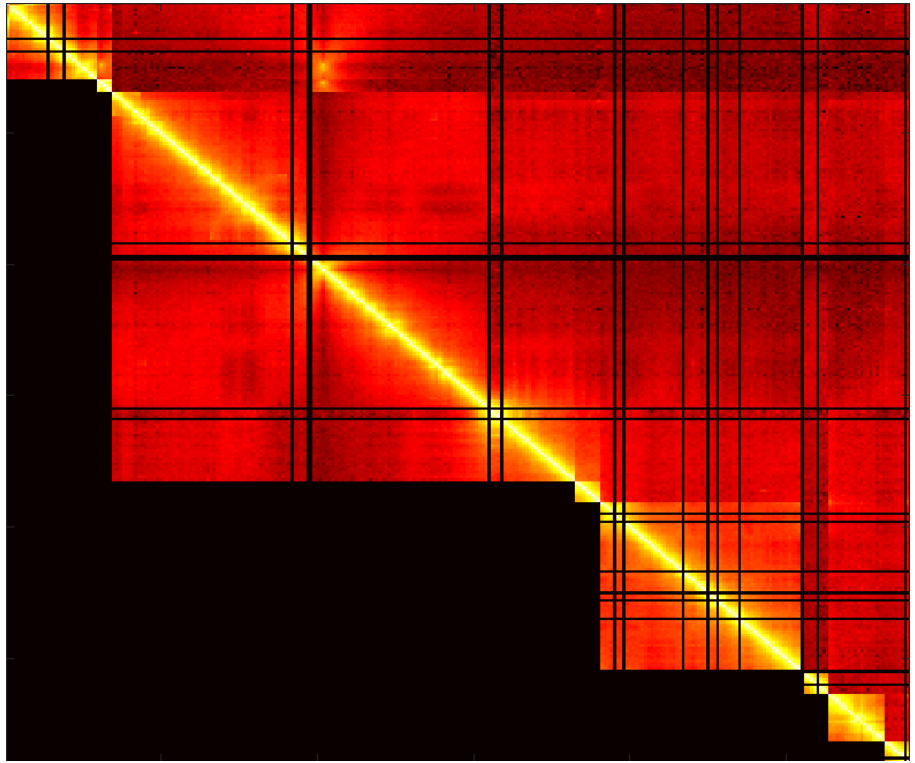




Hi-C guides re-assembly

Many small contigs

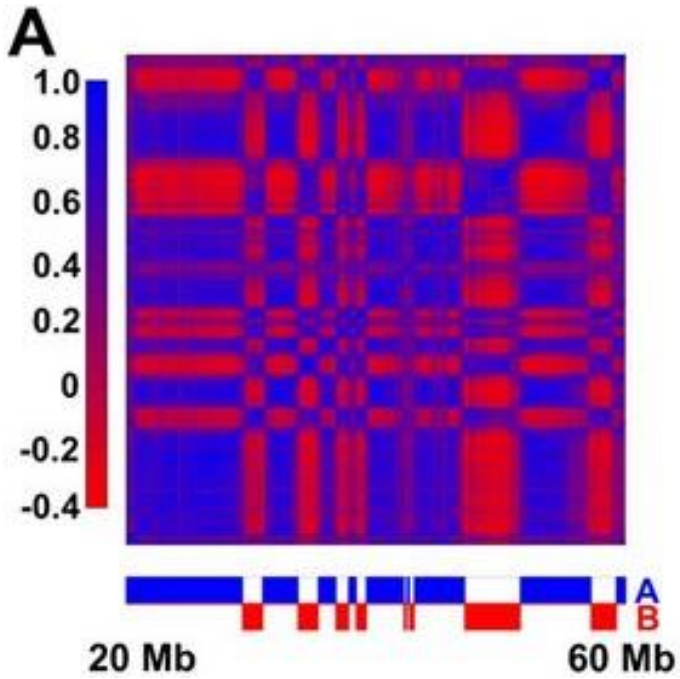


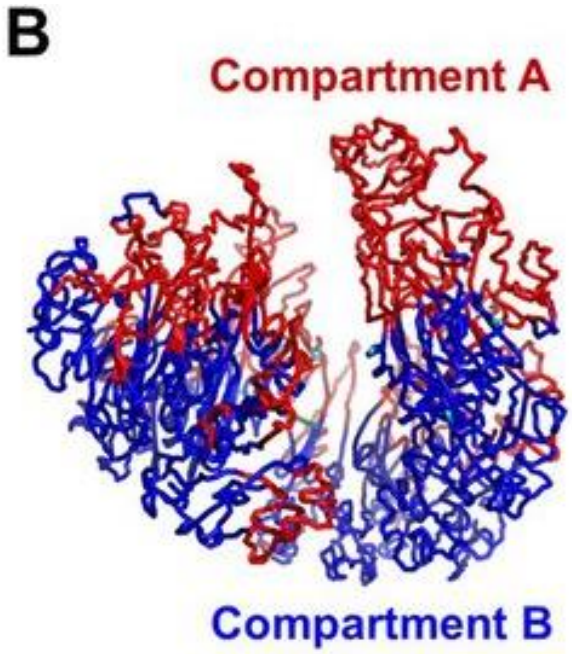


A/B compartment and topologically associating domain (TAD)

A/B compartment

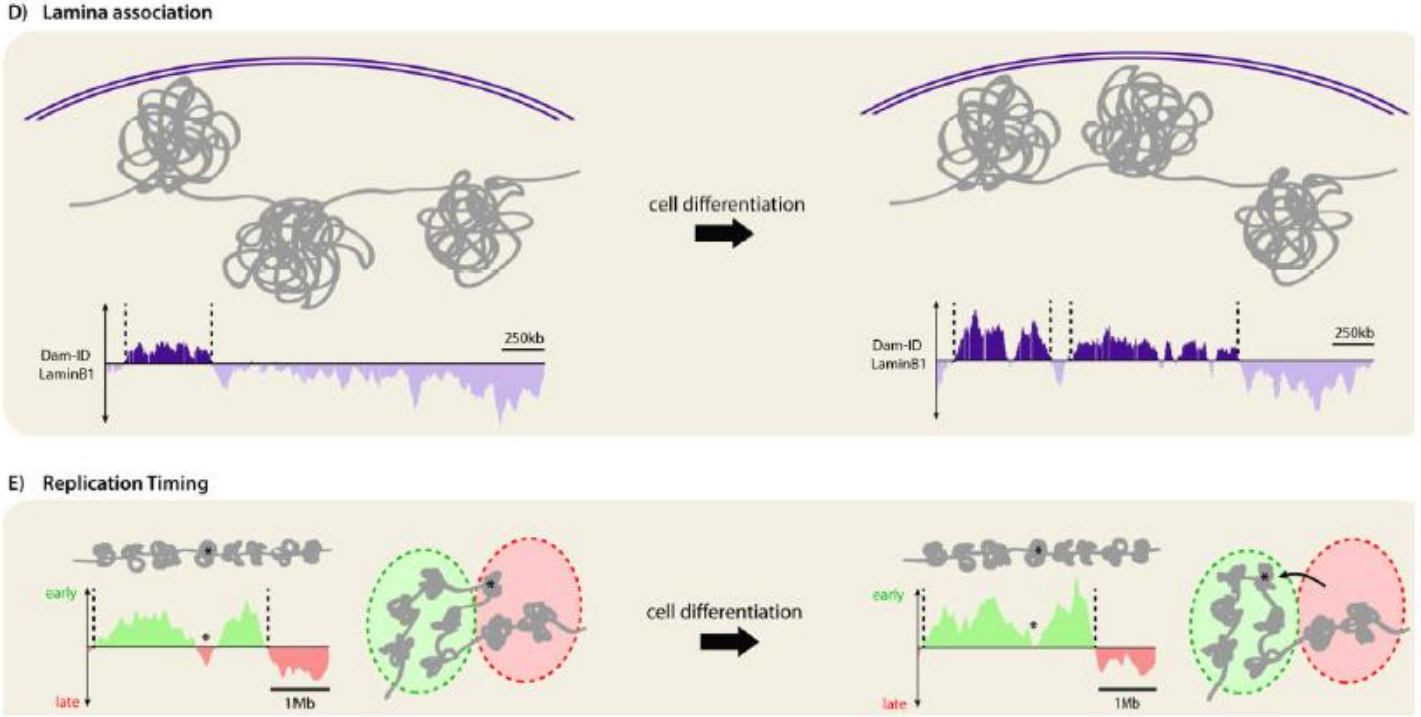
Xie at al. Scientific Reports (2017)





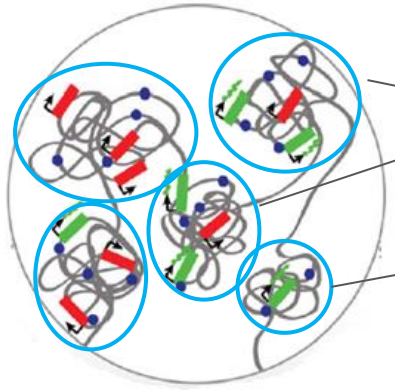
- Caused by localization of euchromatin and heterochromatin

TAD = unit of chromatin structure

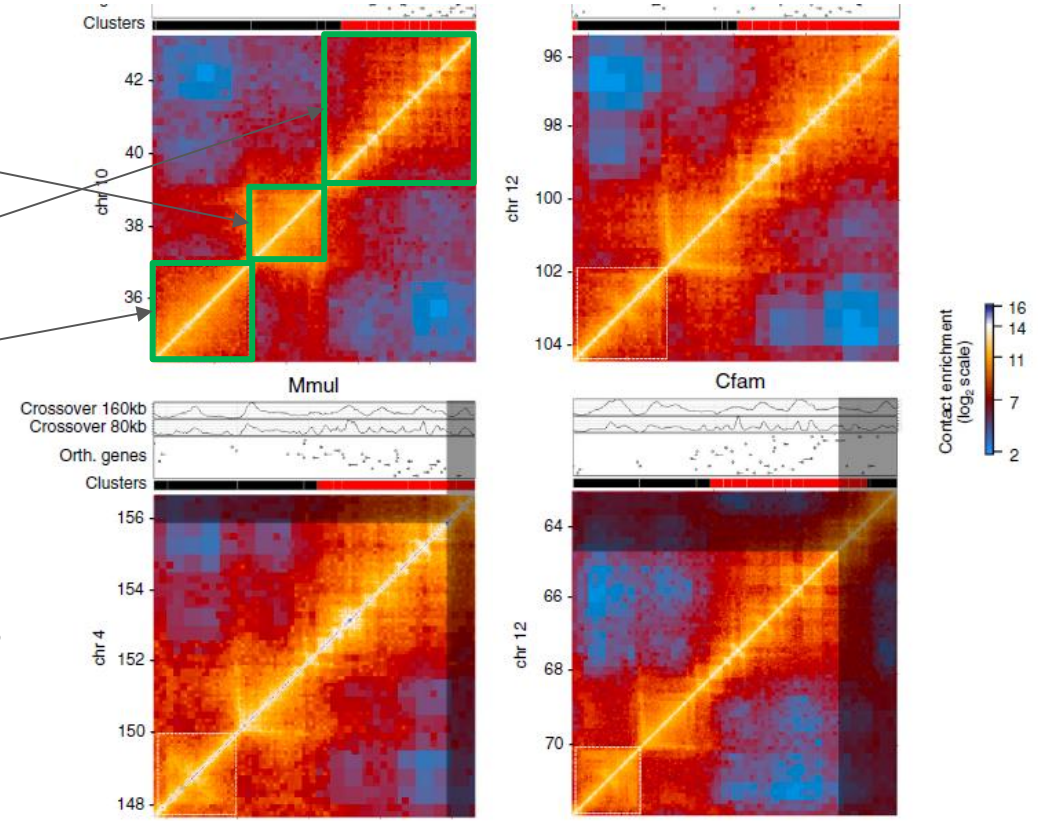


Source: Nora *et al*. Bioessay (2013)

Evolutionary conservation of TAD



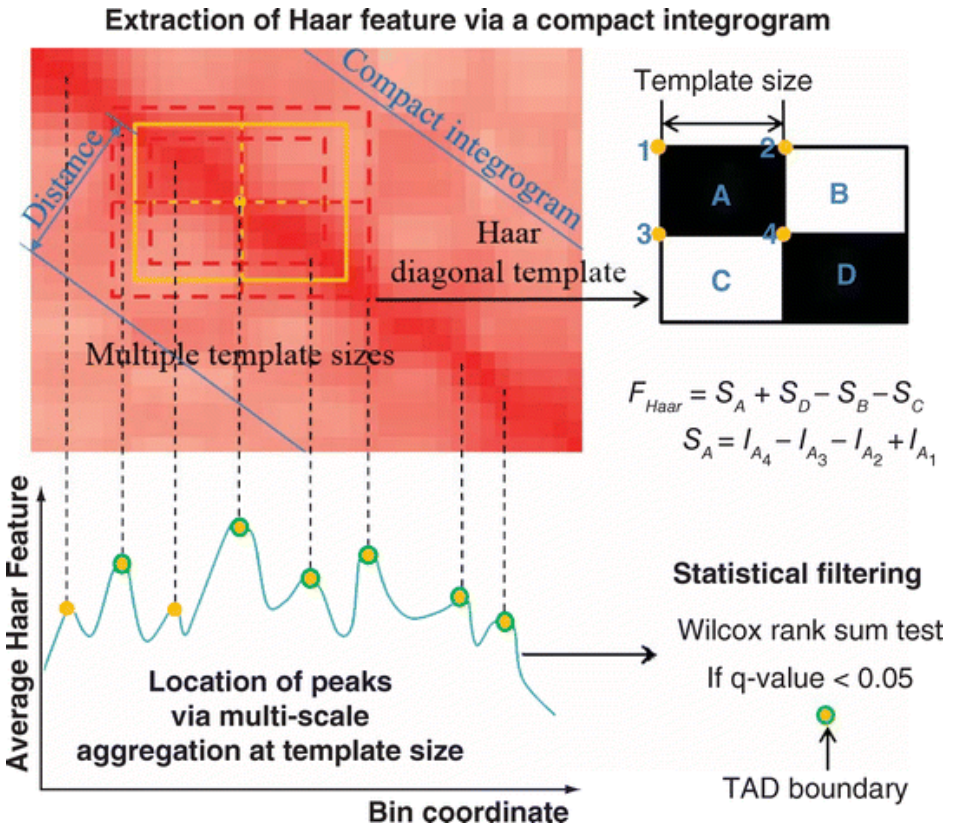
Rudan *et al*. Cell Reports (2015)



- Structural information is (partly) encoded in the DNA sequence

TAD calling

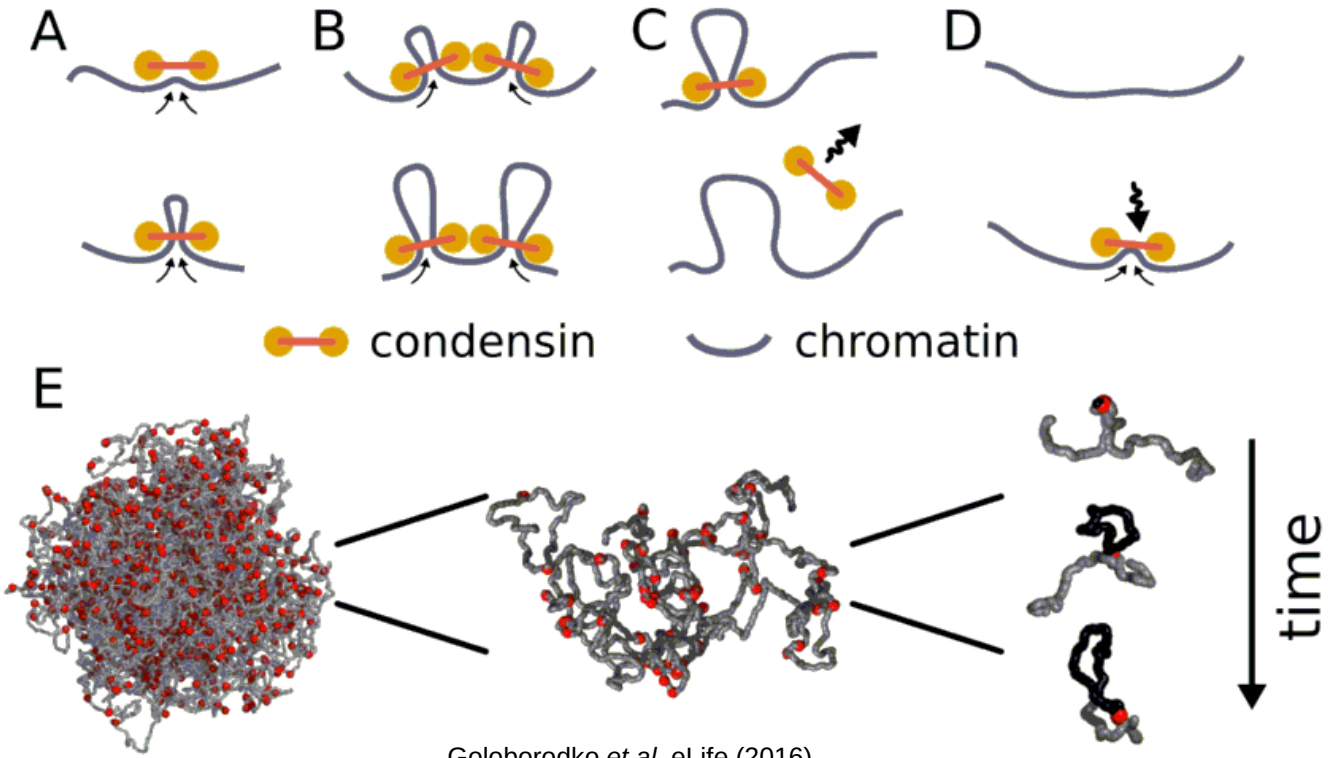
- TAD boundary = location where interaction difference between A+D and B+C is maximized
- Peak detection
- Mann-Whitney U test



Lyu *et al.*, Biotechniques (2020)

Loop extrusion model

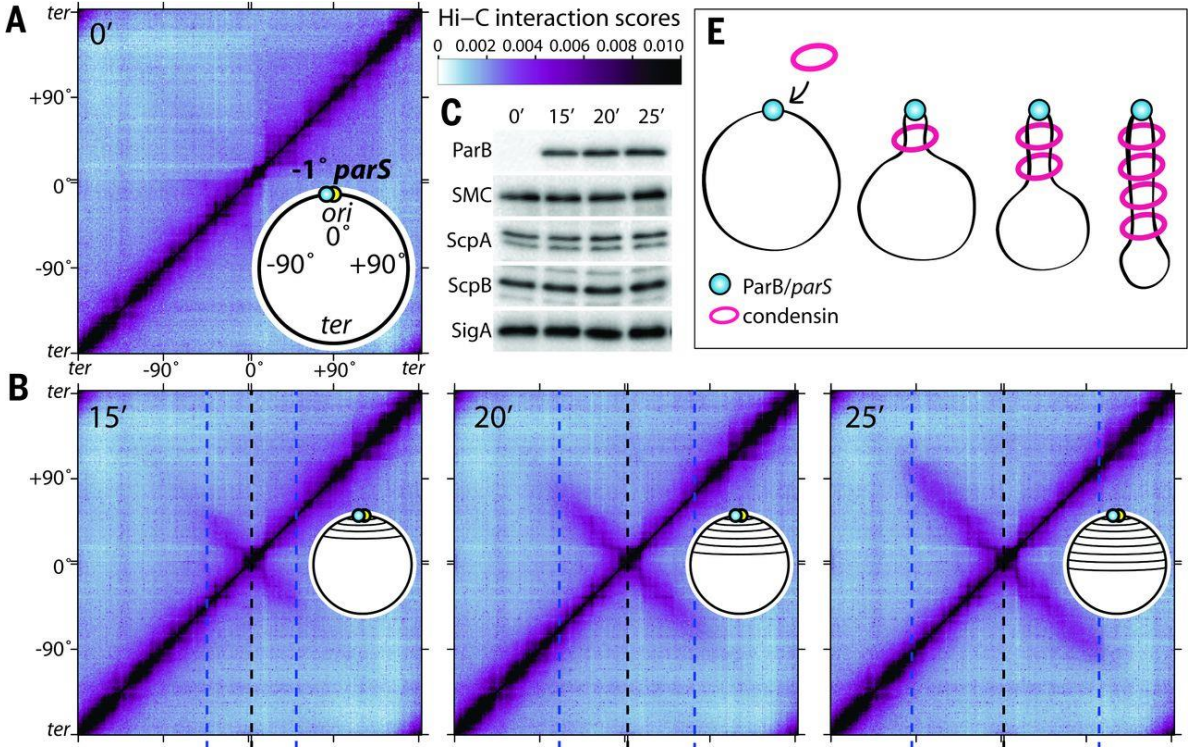
Loop extrusion by proteins



Goloborodko *et al*. eLife (2016)

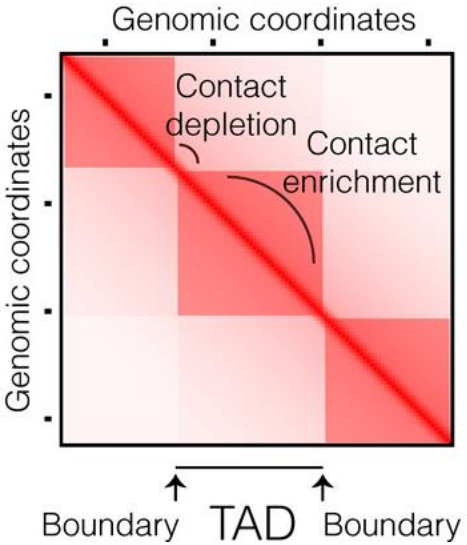
Live loop extrusion in a bacteria

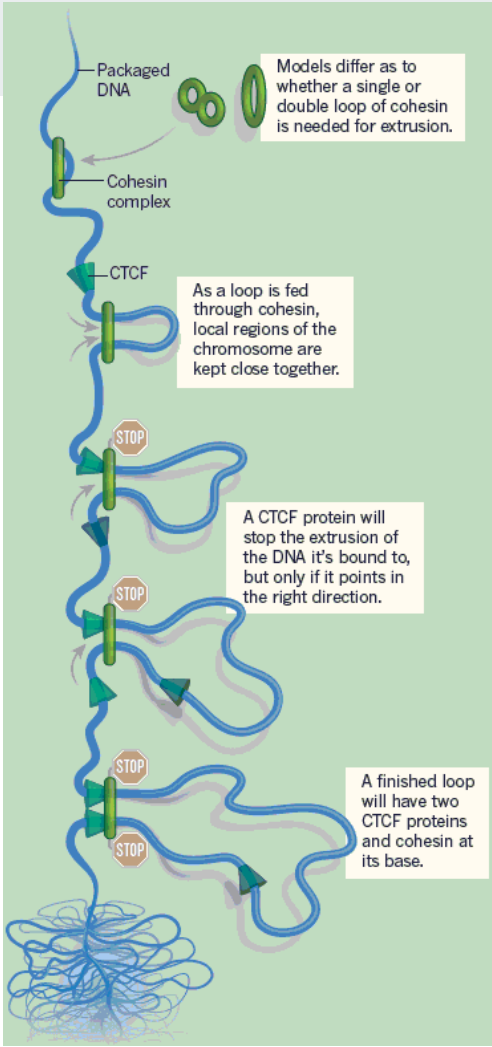
Wang *et al*. Science 355:524-527 (2017)



Loop extrusion and CTCF

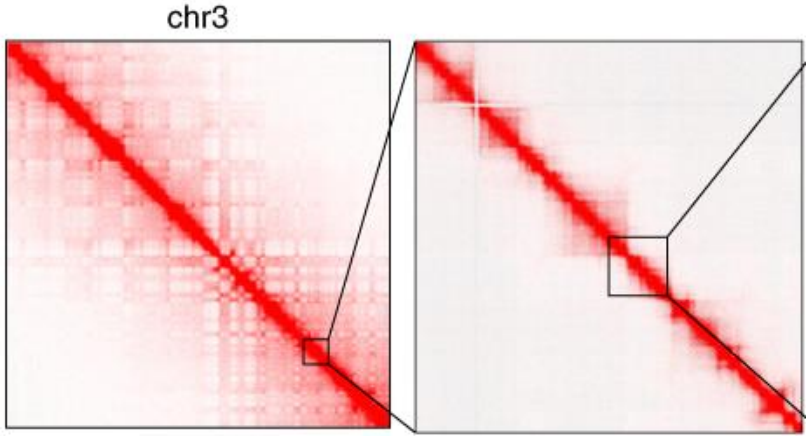
- Proteins like cohesin and condensin extrude loops
- How to stop at TAD boundary?
- CTCF proteins can stop extrusion process in a specific direction
- Not every species has CTCF





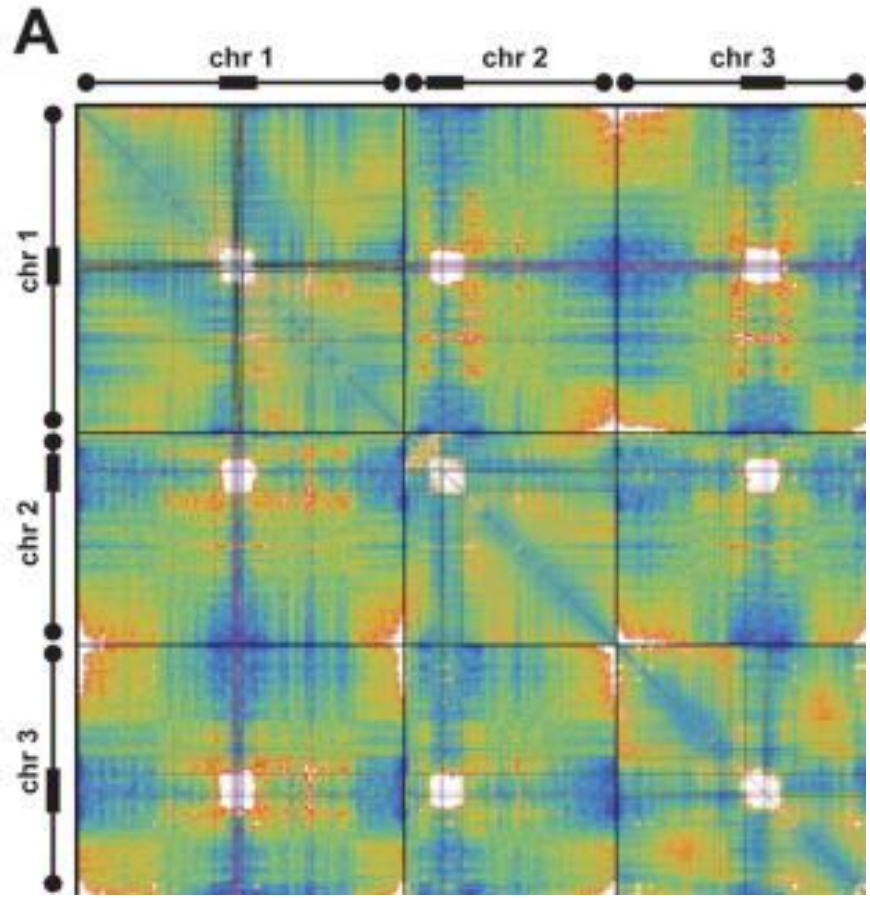
Dolgin. Science (2017)

Human vs arabidopsis



Li et al. Nat Methods 16:991-993 (2019)

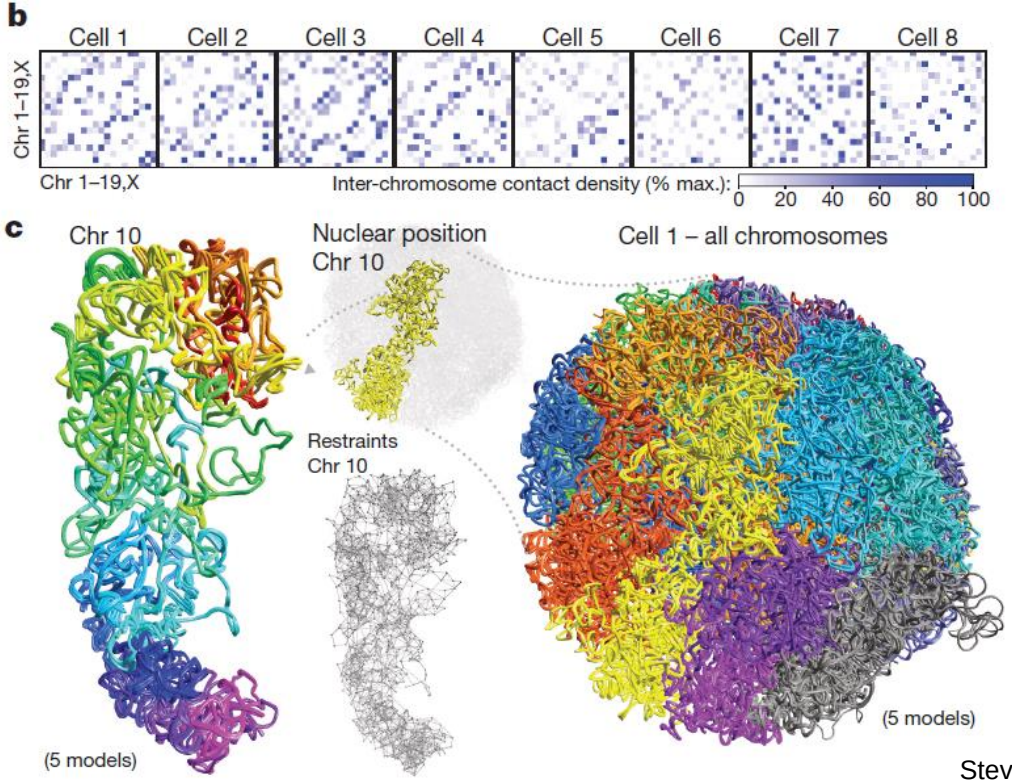
- CTCF is responsible for TADs in human genome



Feng et al. Mol Cell 55:694-707 (2014)

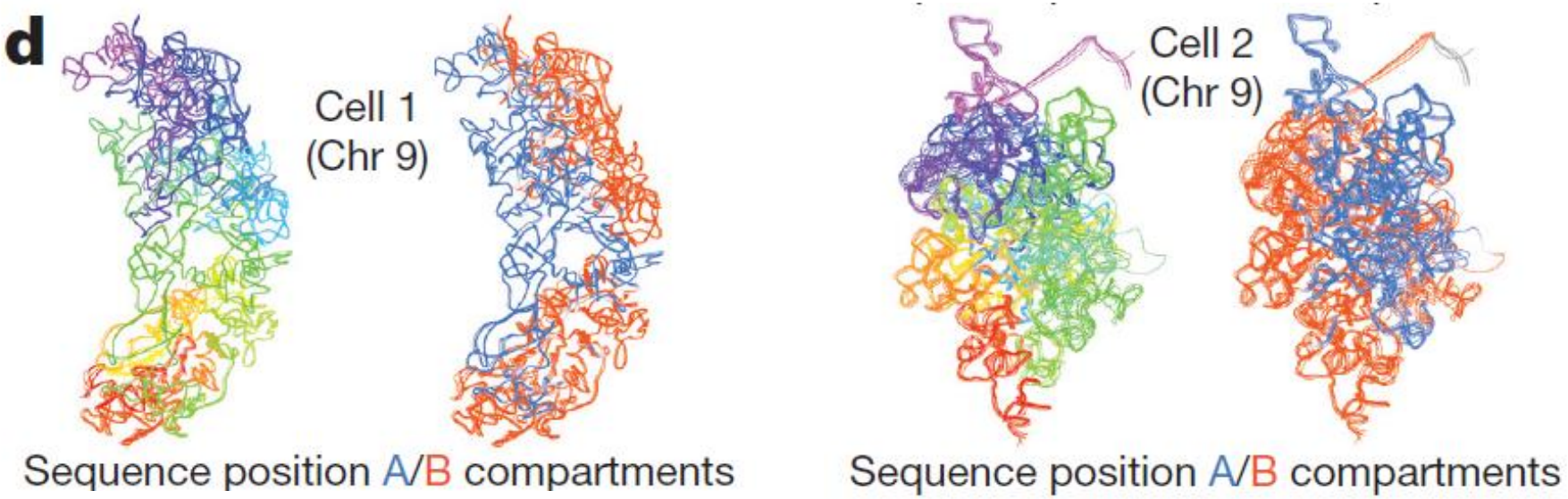
Chromatin structure is not fixed

Single-cell Hi-C



Stevens *et al*. Nature (2017)

Cell-to-cell variations



Any question?