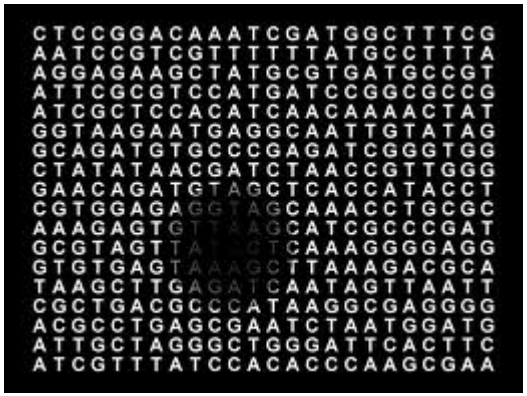
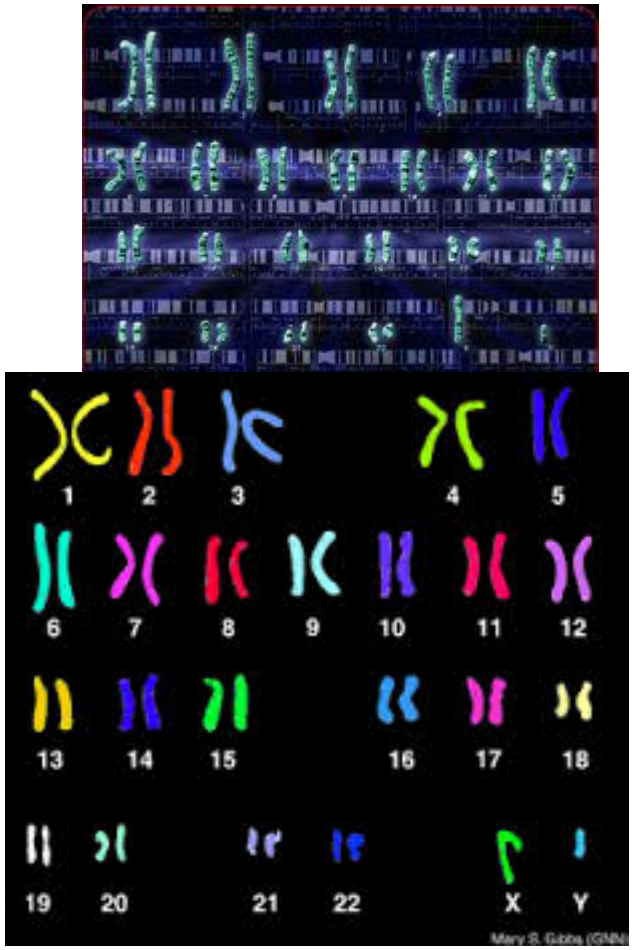
2D and 3D Genome Structure Modeling

Jianlin Cheng, PhD

Associate Professor
Computer Science Department
Informatics Institute
University of Missouri, Columbia
2013

Genome – Code of Life

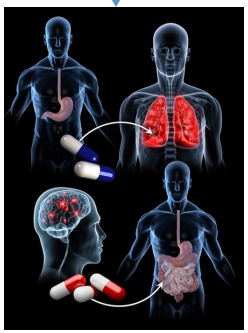




Genome Sequencing (1D)

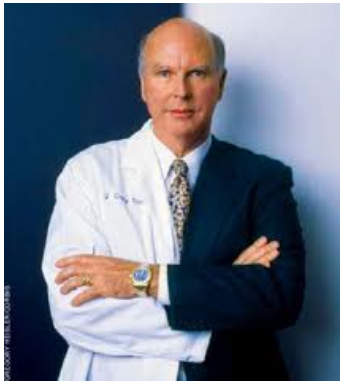
The Genomic Era

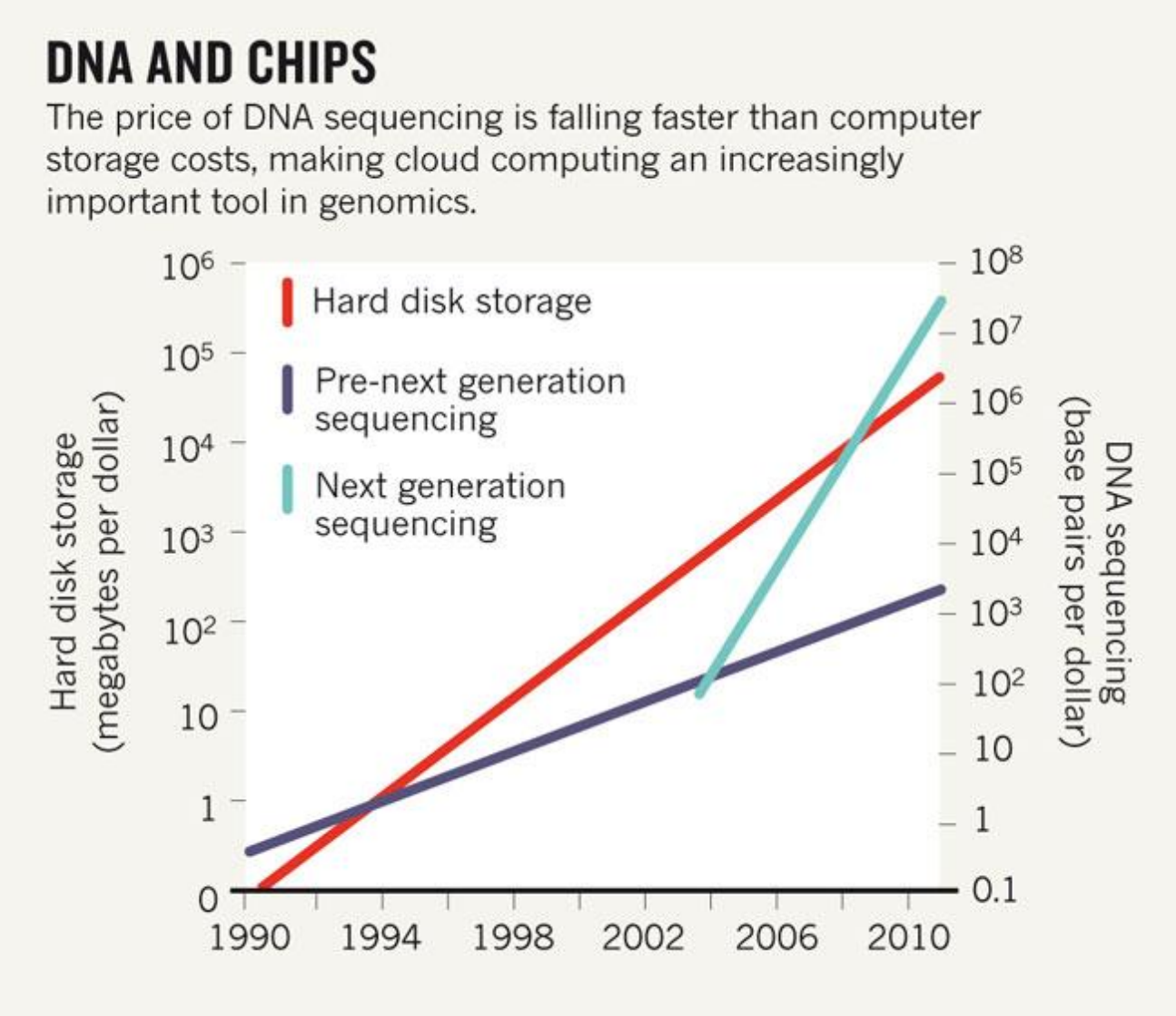
Collins, Venter, Human Genome, 2000



Personal Genome

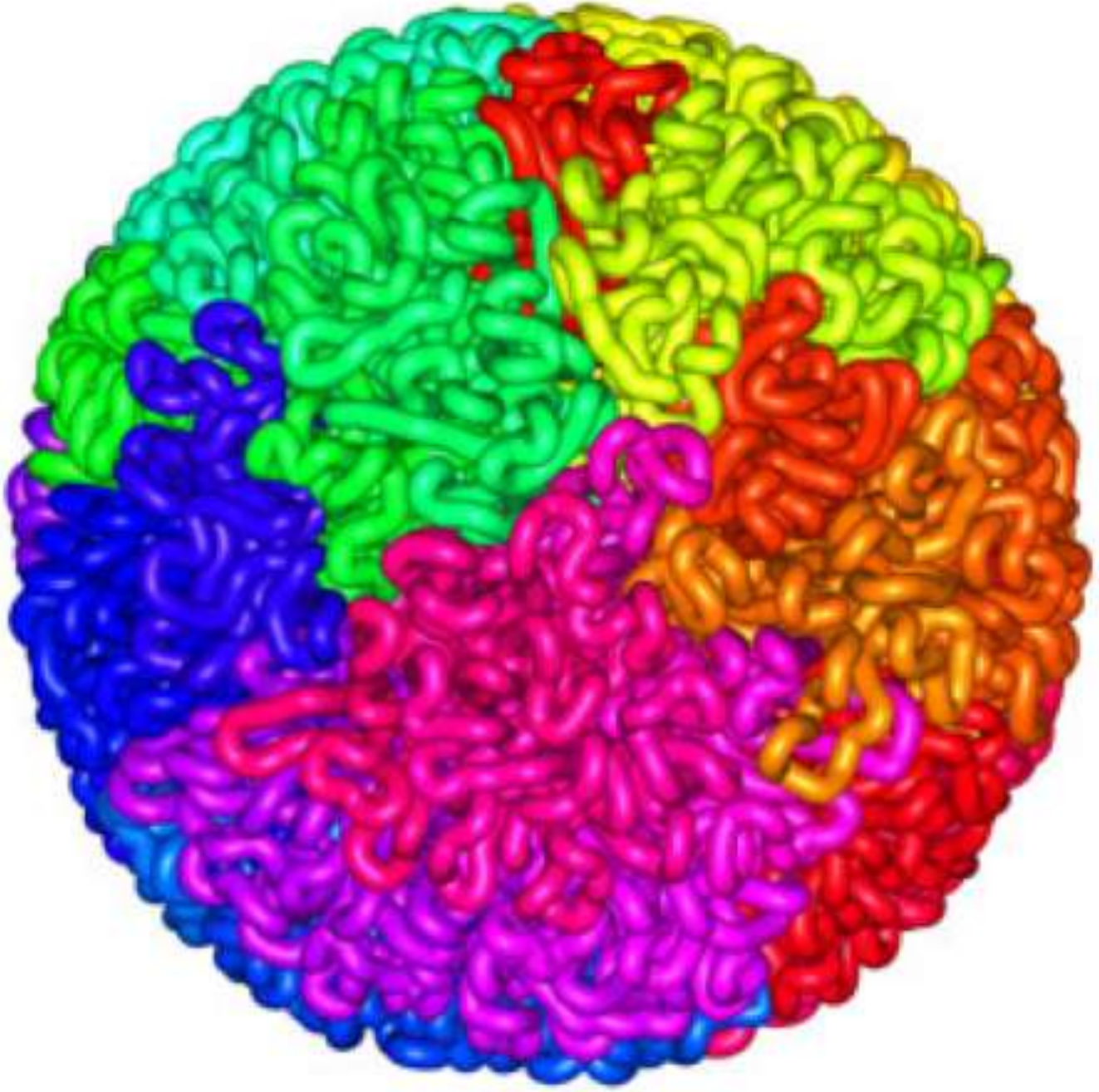
Sequencing Revolution





- **$1000 Personal Genome in 2010s**

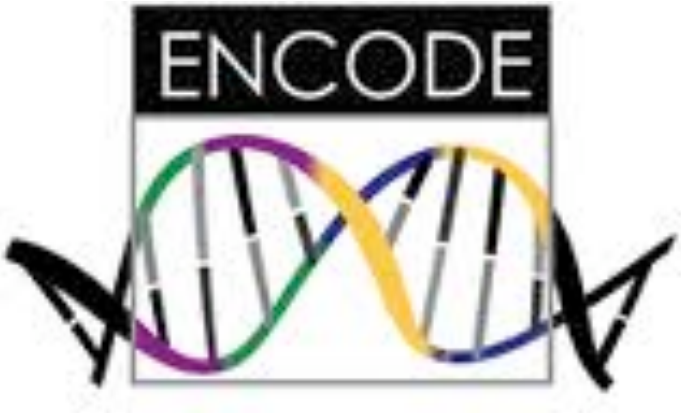
3D Shape



http://images.google.edu

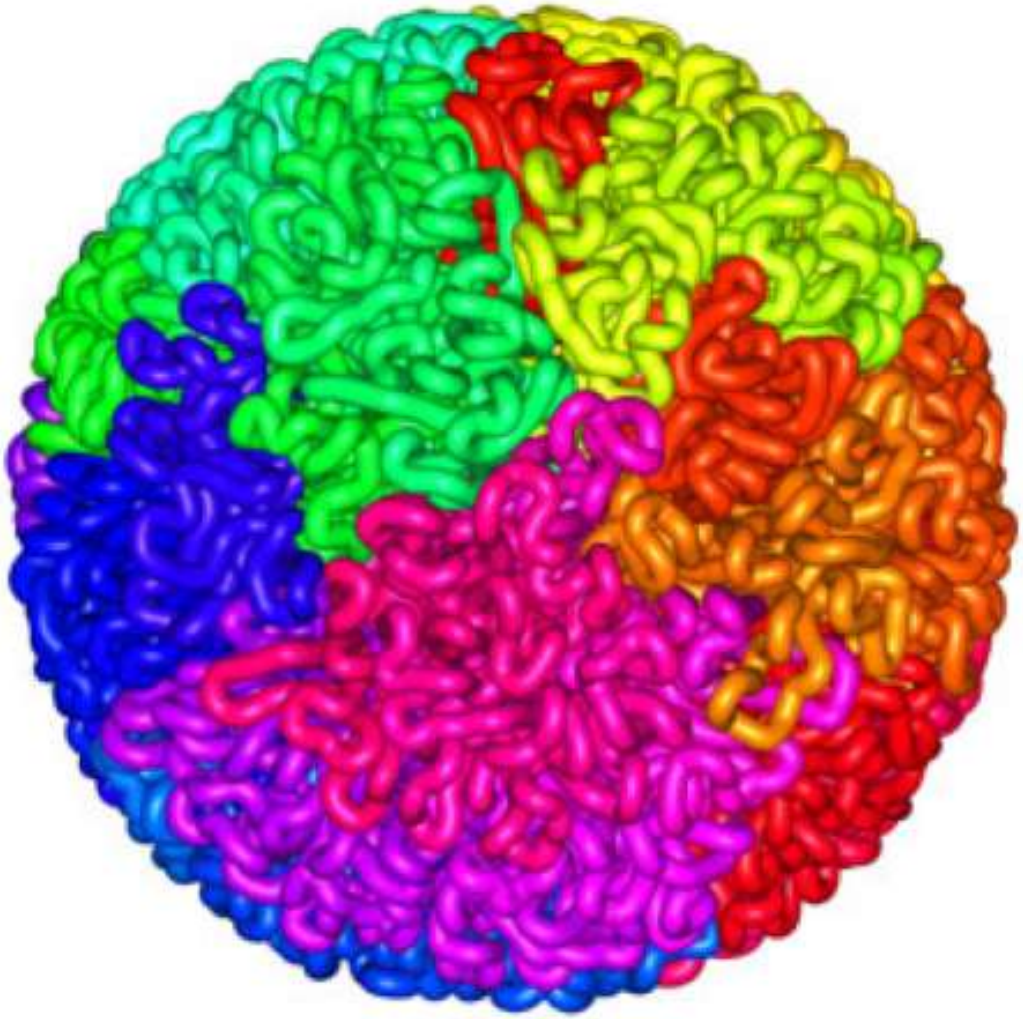
Genome

ATGTATCCAGGTAGTGGACGTTACACCTACAACAACGCTGGTGGTAATAATGGCTACCAA
CGGCCCATGGCTCCTCCACCTAACCAGCAGTATGGACAGCAATATGGTCAGCAATATGAA
CAGCAGTATGGACAGCAATATGGGCAACAAAATGATCAGCAATTCAGTCAGCAATATGCT
CCACCACCAGGTCCTCCCCCTATGGCTTATAACAGGCCTGTGTATCCCCCCCCTCAATTC
CAGCAGGAACAGGCAAAGGCACAATTAAGCAACGGCTACAACAATCCTAATGTAAACGCA
TCCAATATGTACGGTCCACCCCAGAATATGTCATTACCTCCACCTCAAACACAAACTATT
CAAGGTACAGACCAACCTTATCAGTATTCTCAATGTACTGGGCGTAGAAAGGCTTTGATT
ATCGGTATAAACTACATAGGTTCAAAAAATCAACTGCGTGGTTGTATCAATGATGCTCAT
AACATCTTCAACTTTTTGACTAATGGGTACGGTTACAGTTCAGATGACATTGTCATATTA
ACTGATGATCAGAACGATTTGGTCAGGGTTCCCACTAGGGCTAATATGATTAGGGCCATG
CAATGGTTGGTCAAGGATGCGCAACCCAATGATTCTTTGTTCCTTCATTATTCTGGACAT
GGTGGCCAAACTGAAGATTTGGATGGGGACGAAGAAGATGGGATGGATGATGTTTATATAT
CCGGTCGATTTCGAAACTCAAGGGCCAATTATCGACGATGAAATGCACGATATAATGGTG
AAGCCCTTACAACAAGGTGTTAGACTAACAGCATTGTTTGACTCTTGTCATTCGGGTACA
GTGTTGGATCTTCCATATACCTATTCTACTAAGGGTATTATTAAGGAGCCCAATATTTGG
AAGGATGTTGGCCAAGATGGCCTGCAAGCAGCTATTTCATATGCCACAGGAAACAGGGCT
GCTTTGATTGGTTCTTTAGGTTCTATATTCAAGACCGTTAAGGGAGGTATGGGCAATAAT
GTGGATAGAGAACGCGTGAGACAGATCAAATTCTCAGCAGCAGATGTTGTTATGTTATCA
GGTTCGAAGGATAATCAAACTTCTGCAGATGCTGTCGAAGATGGGCAAAATACAGGTGCA
ATGTCCCACGCCTTCATCAAGGTTATGACTTTACAACCACAGCAATCATATTTATCTCTT
TTACAGAACATGAGGAAAGAATTGGCTGGTAAGTATTCTCAAAAACCACAATTATCATCG
TCACACCCTATTGACGTAAATCTGCAATTTATTATGTAG



>95% non-coding regions of a genome are not junk!!

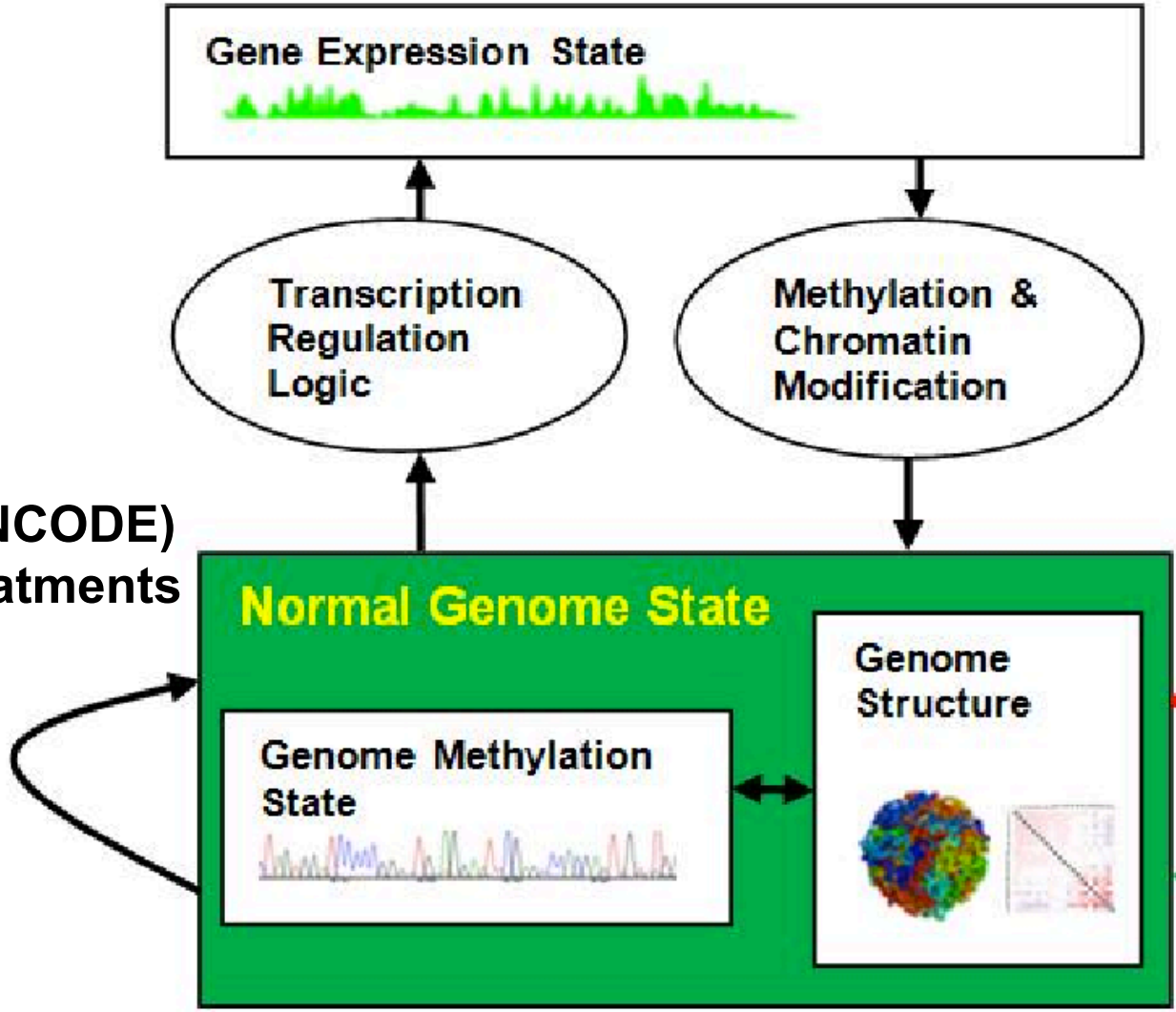
Genome Conformation



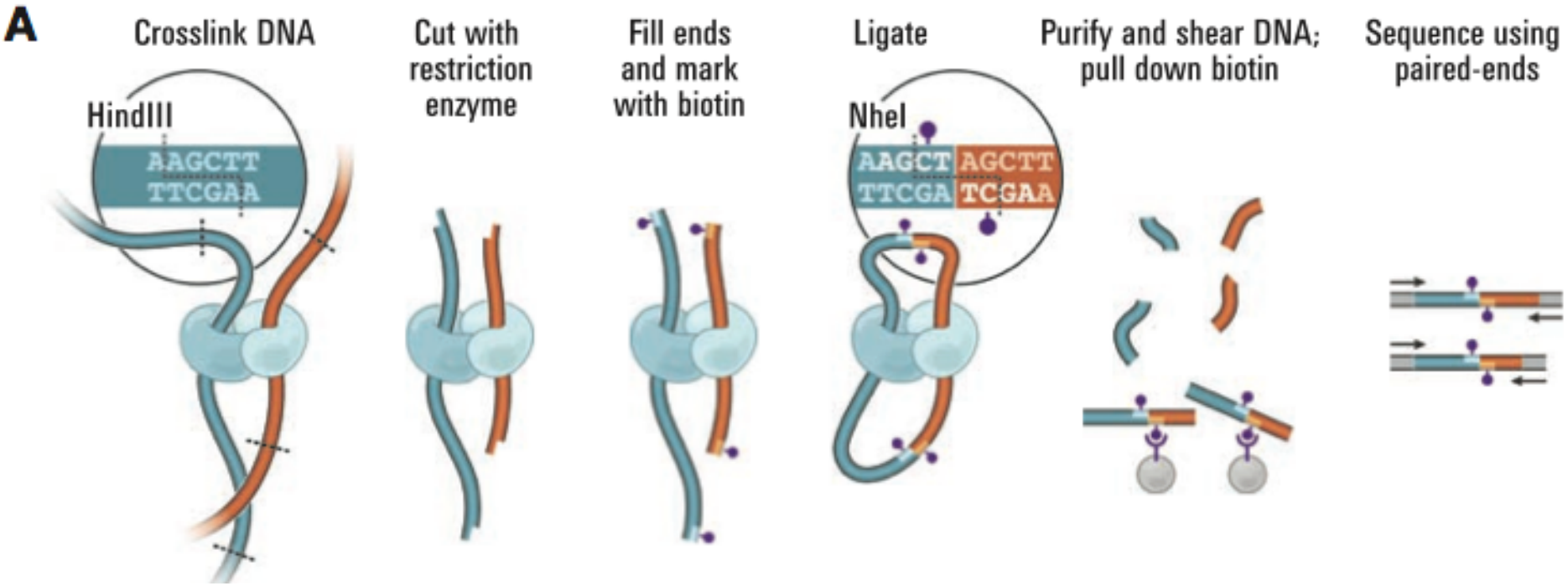
Liberman-Aiden et al., 2009

3D Genome Structure is Important

- Spatial gene regulation
- Transcription efficiency
- Genome interpretation
- Function implication (ENCODE)
- Disease diagnosis & treatments
- Drug design



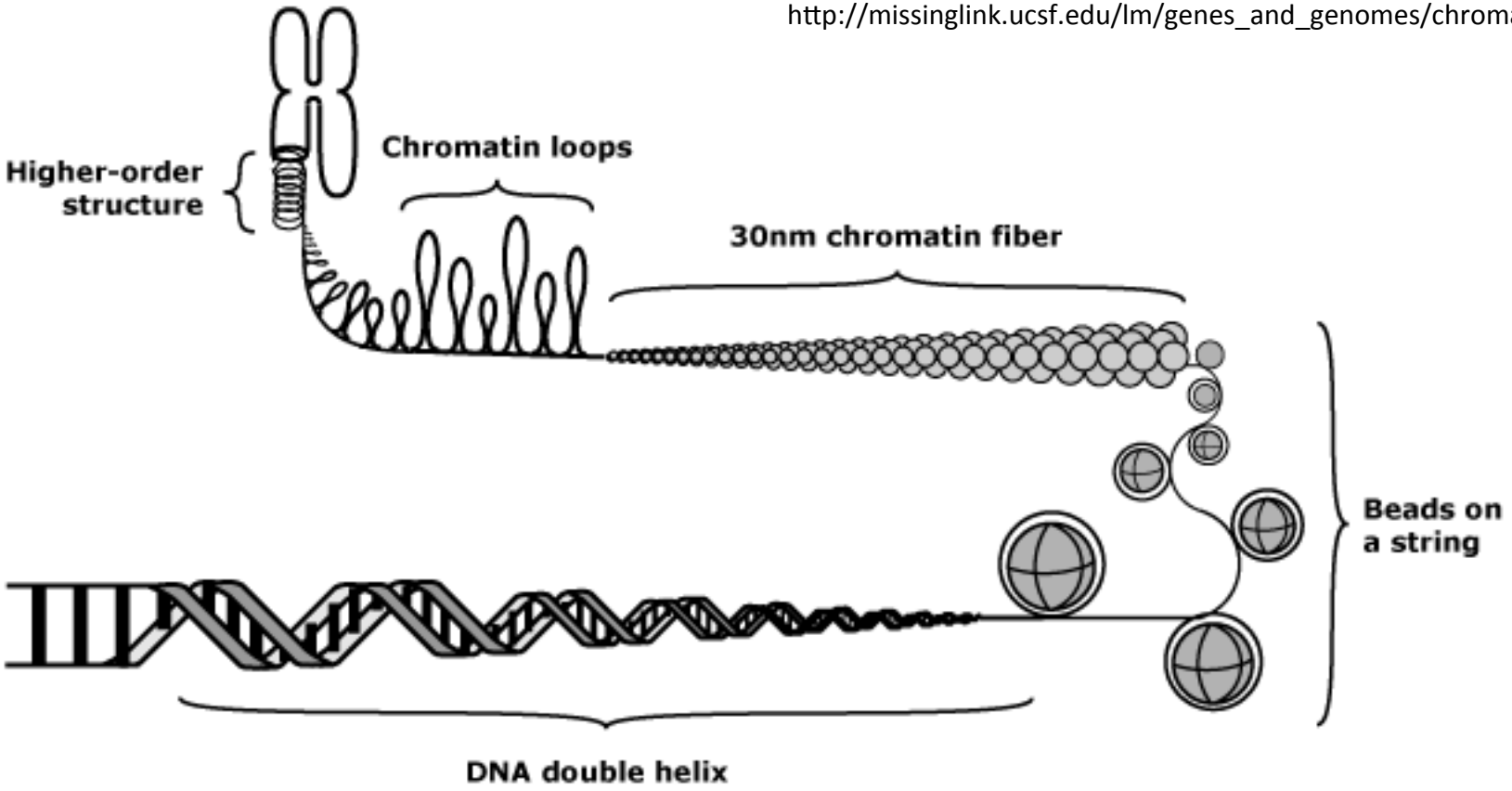
Chromosome Conformation Capturing (Hi-C)



Liberman-Aiden et al., 2009

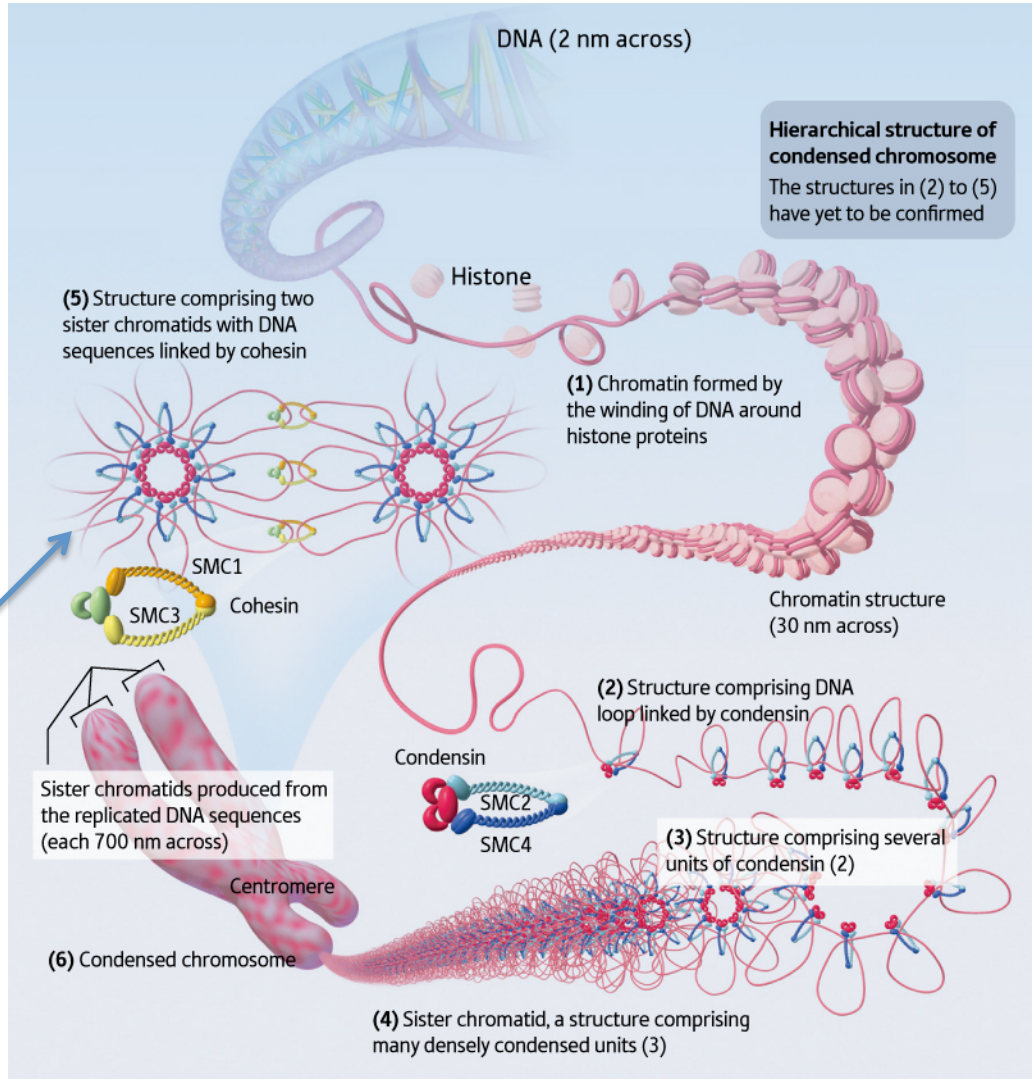
Multi-Level Chromosome Structure

http://missinglink.ucsf.edu/lm/genes_and_genomes/chromatin.html



DNA is further packaged. **Nucleosomes** are arranged together into **a fiber** approximately 30 nanometers in diameter. The *precise structure of the chromatin fiber is not known*. Chromatin fiber is further organized into **chromatin loops**, and chromatin loops are further organized into higher-order structures. It has been suggested that **packaging plays a role in gene expression** (gene expression may require associated DNA to open up and acquire an unpackaged conformation). The fully condensed chromosome structure is only seen during mitosis.

Model 1 (Riken)

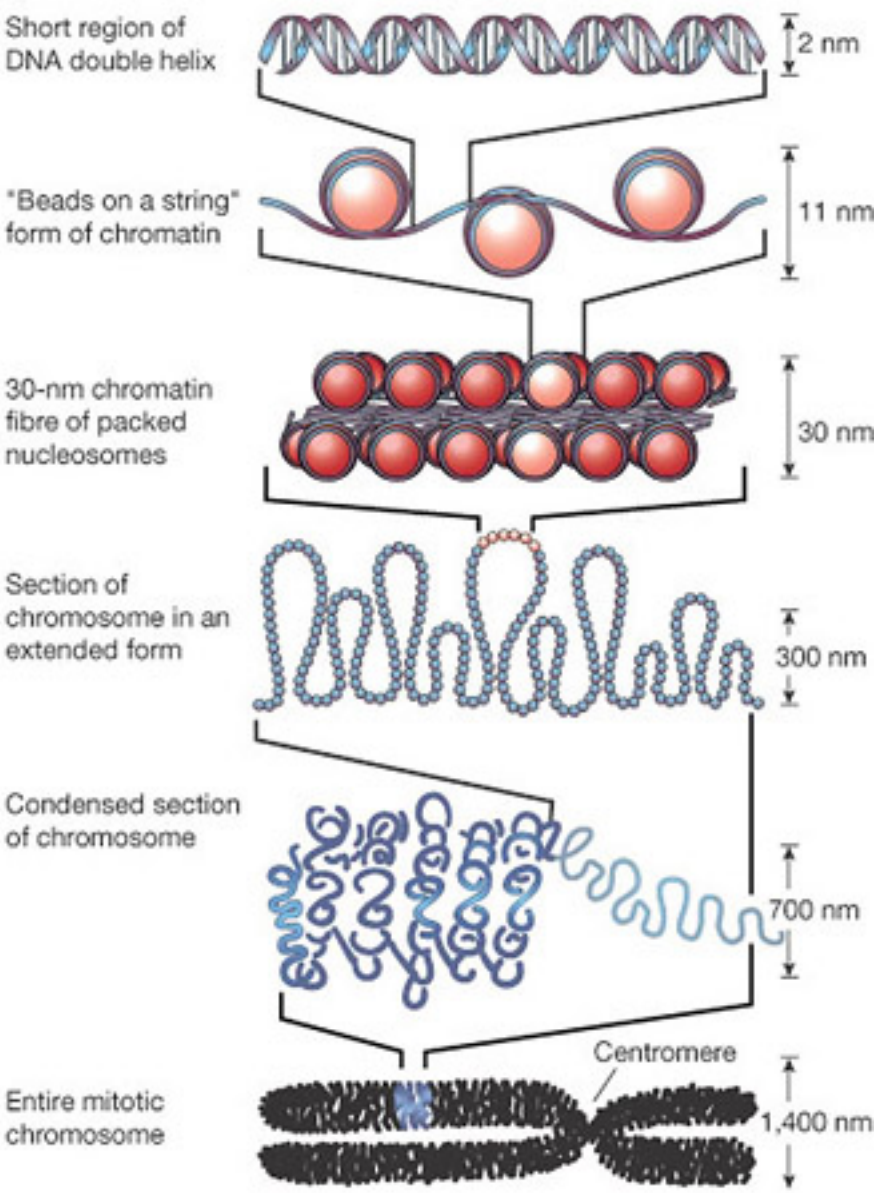


Globular Structure

The DNA is a remarkable molecule in many ways. It encodes our entire genome, and if stretched out in a thin thread would measure **1.8 m** in length.

http://www.rikenresearch.riken.jp/eng/frontline/6485

Size of Elements

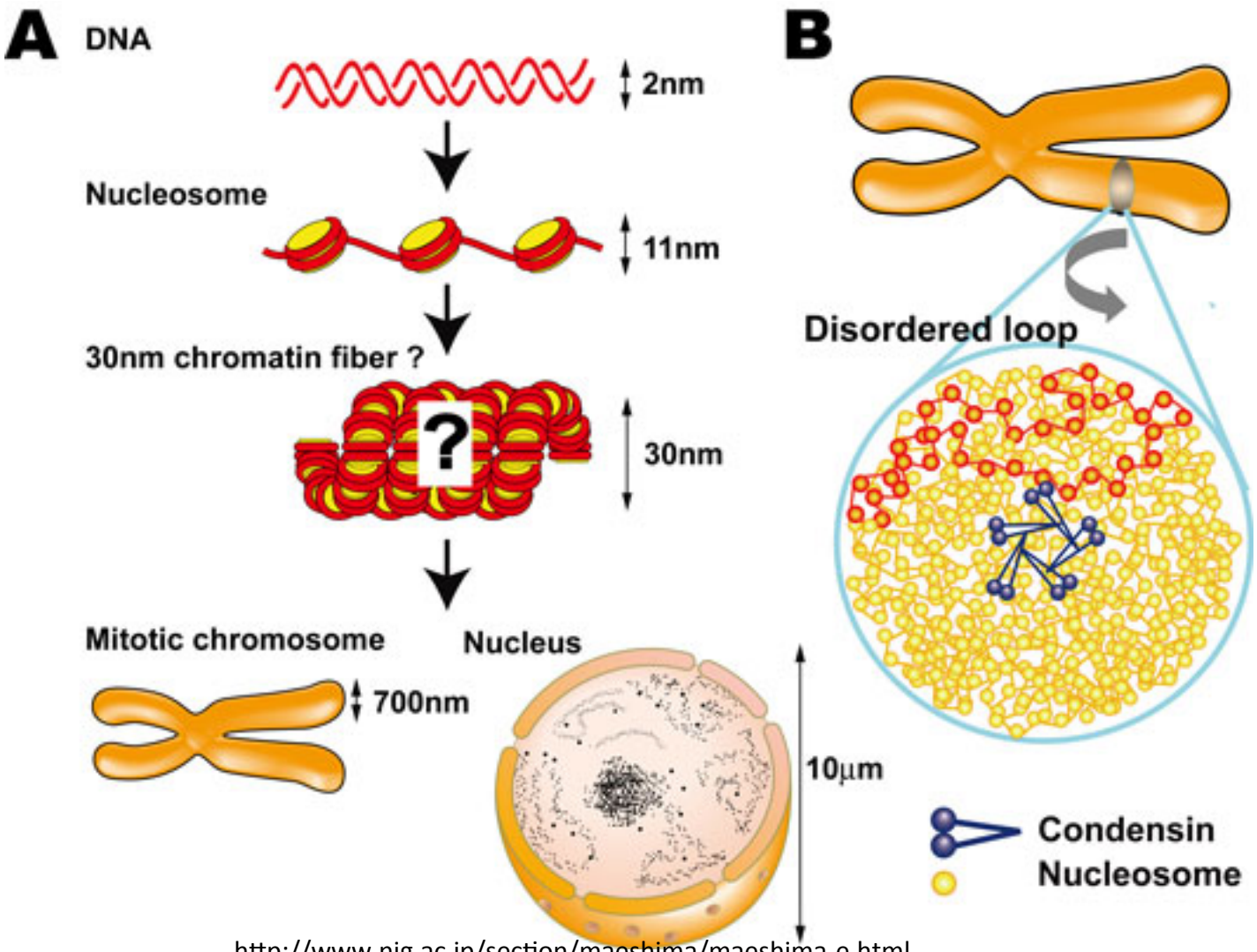


NM: 10^{-9} meter

A complex of DNA and basic proteins (such as **histone**) in eukaryotic cells that is condensed into chromosomes in mitosis and meiosis. There are two types. **Heterochromatic** is densely-coiled chromatin that appears as nodules in or along chromosomes and contains relatively few genes. **Euchromatic** is the less-coiled and genetically active portion of chromatin that is largely composed of genes.

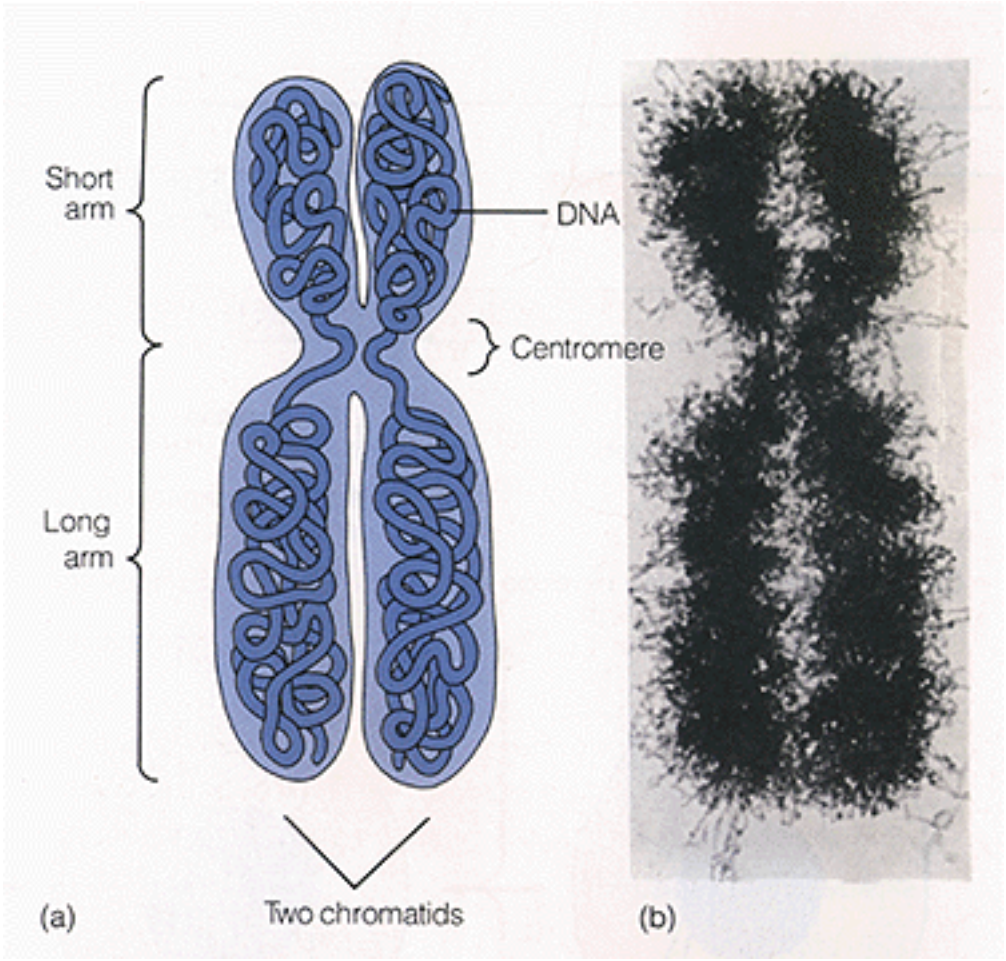
http://www.daviddarling.info/encyclopedia/C/chromatin.html

Size of Elements



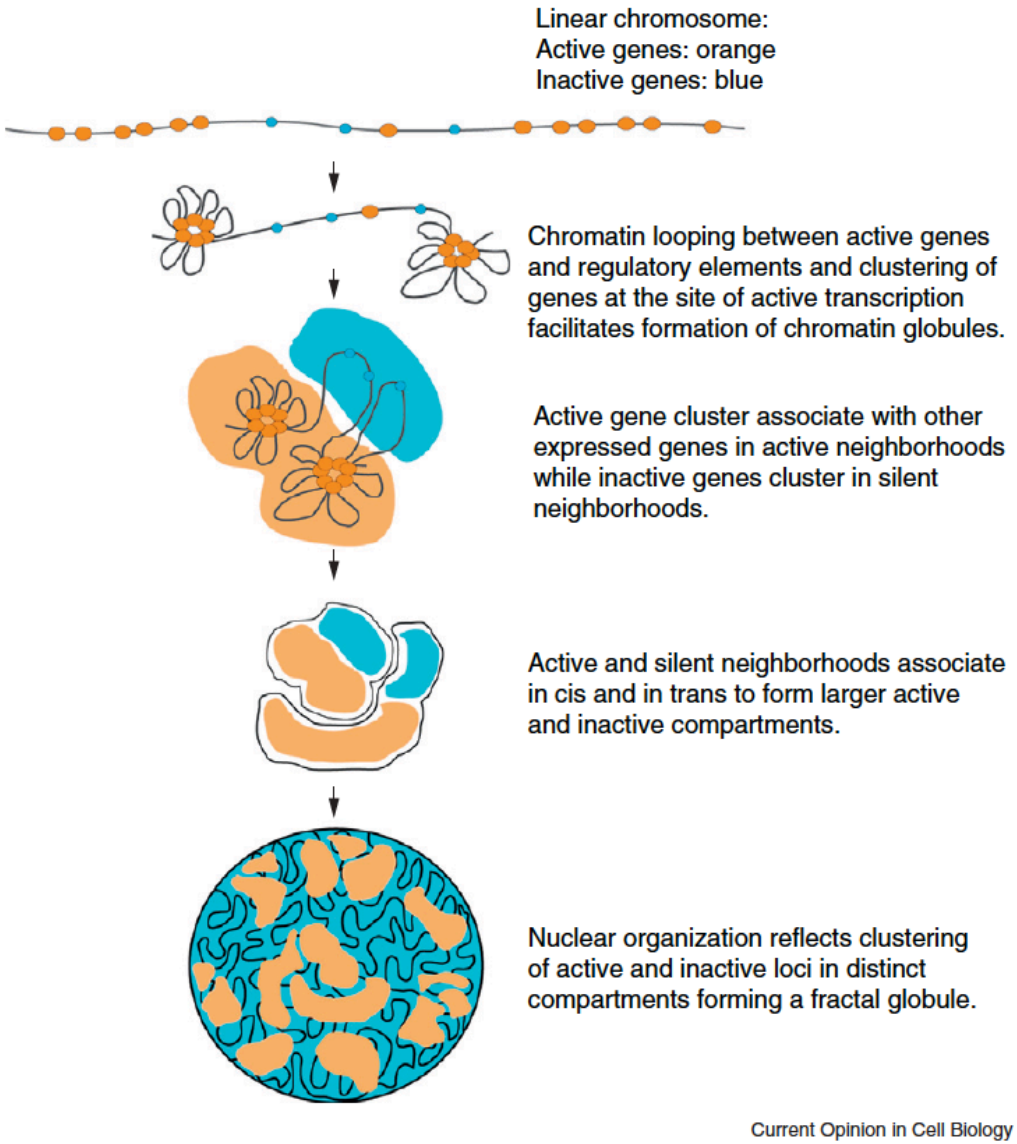
http://www.nig.ac.jp/section/maeshima/maeshima-e.html

Chromosomes



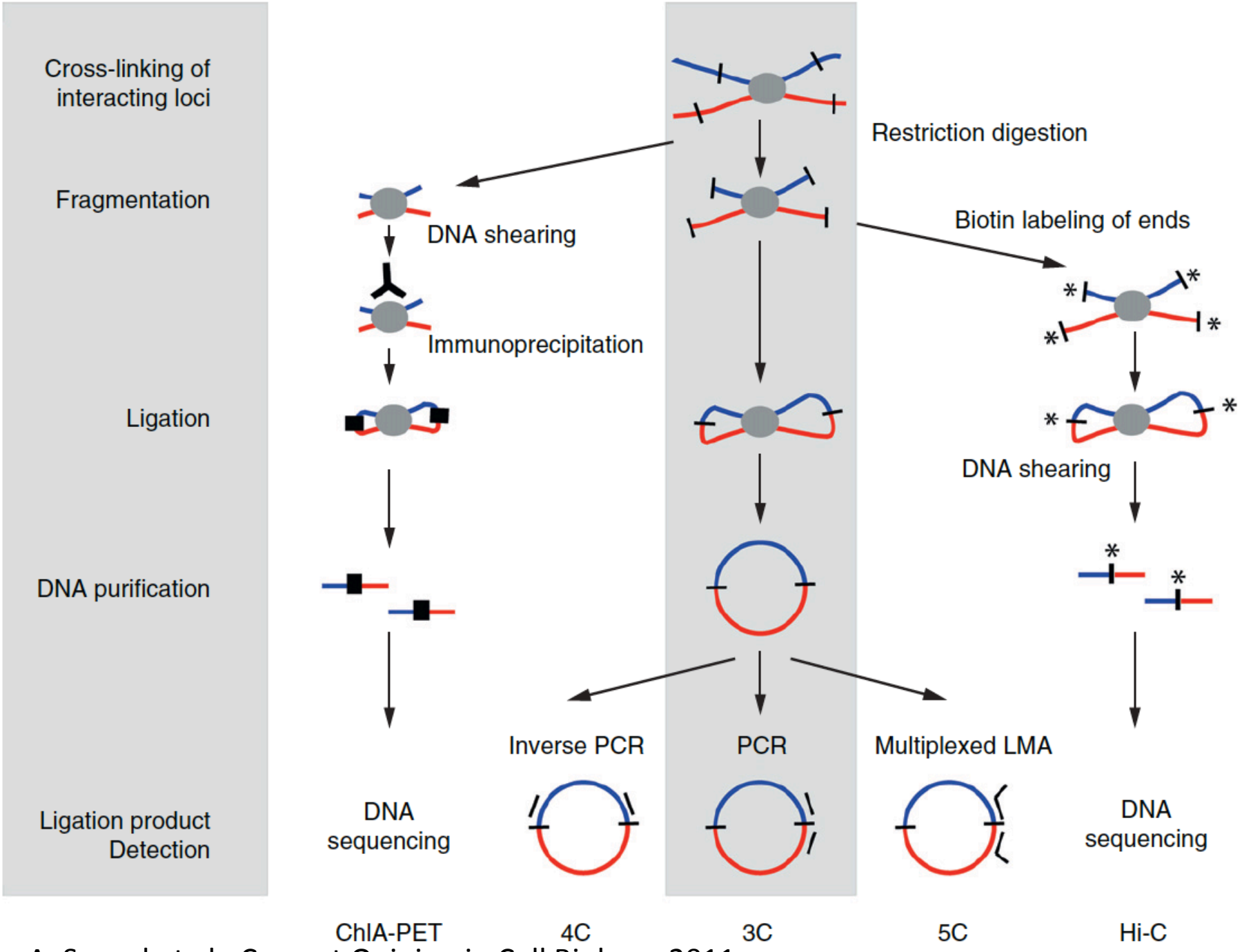
http://www.copernicusproject.ucr.edu/ssi/HSBiologyResources.htm

Two Compartments



A. Sanyal et al., Current Opinion in Cell Biology, 2011

Chromosome Conformation Capturing Techniques

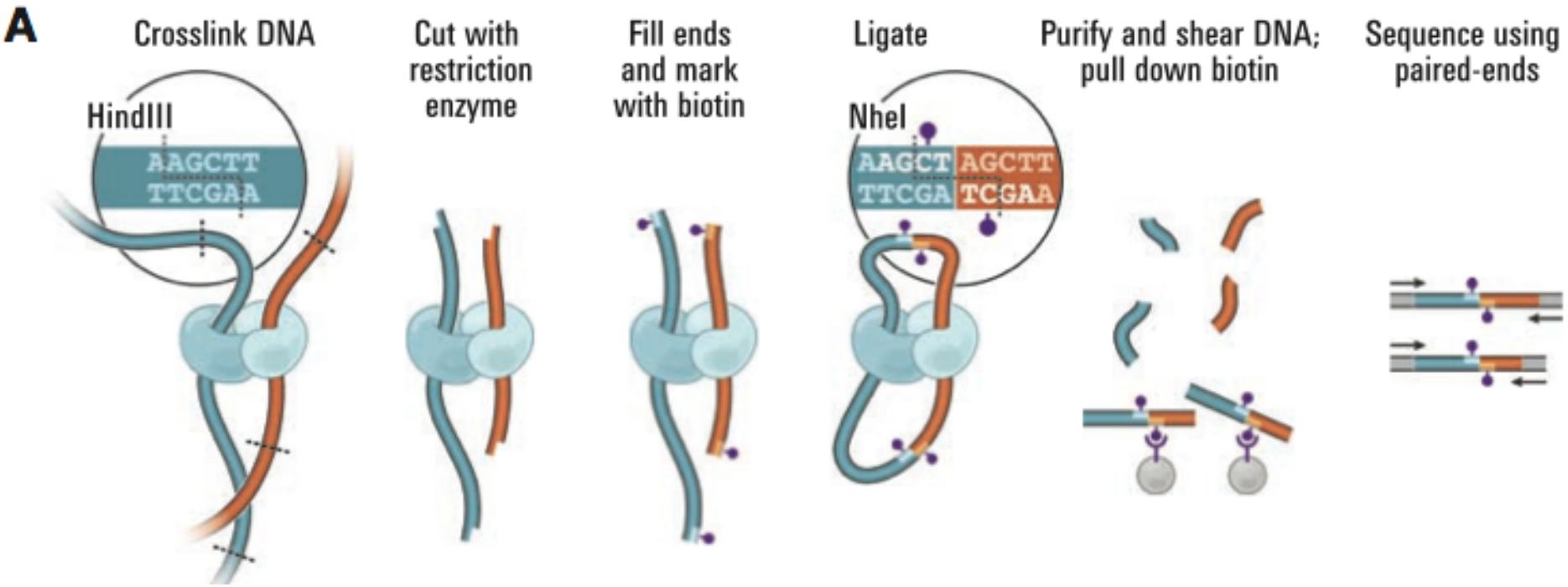


A. Sanyal et al., Current Opinion in Cell Biology, 2011

3C, 4C, 5C, Hi-C Methods

- All 3C-based methods use the same principle of chromatin interaction detection (indicated in grey): *cross-linking of interacting loci with formaldehyde, followed by DNA fragmentation, DNA ligation, DNA purification and finally ligation product detection.*
- 3C employs restriction enzymes to fragment DNA, and regular PCR to detect ligation products, while 4C employs inverse PCR to detect all fragments ligated to a locus of choice.
- 5C uses multiplexed ligation mediated amplification (LMA) to detect large numbers of interactions simultaneously using pools of primers for
- thousands of loci of interest.
- ChIA-PET employs sonication to fragment cross-linked chromatin followed by an immunoprecipitation step before DNA ligation to enrich for loci bound by a protein of interest. Linkers are then ligated (black thick lines) and DNA is analyzed by direct deep sequencing.
- Hi-C employs restriction enzymes to fragment chromatin followed by filling in of the staggered ends using biotinylated nucleotides before DNA ligation. DNA is sheared and DNA fragments containing ligation junctions are purified using streptavidin-coated beads. DNA is then directly deep sequenced.

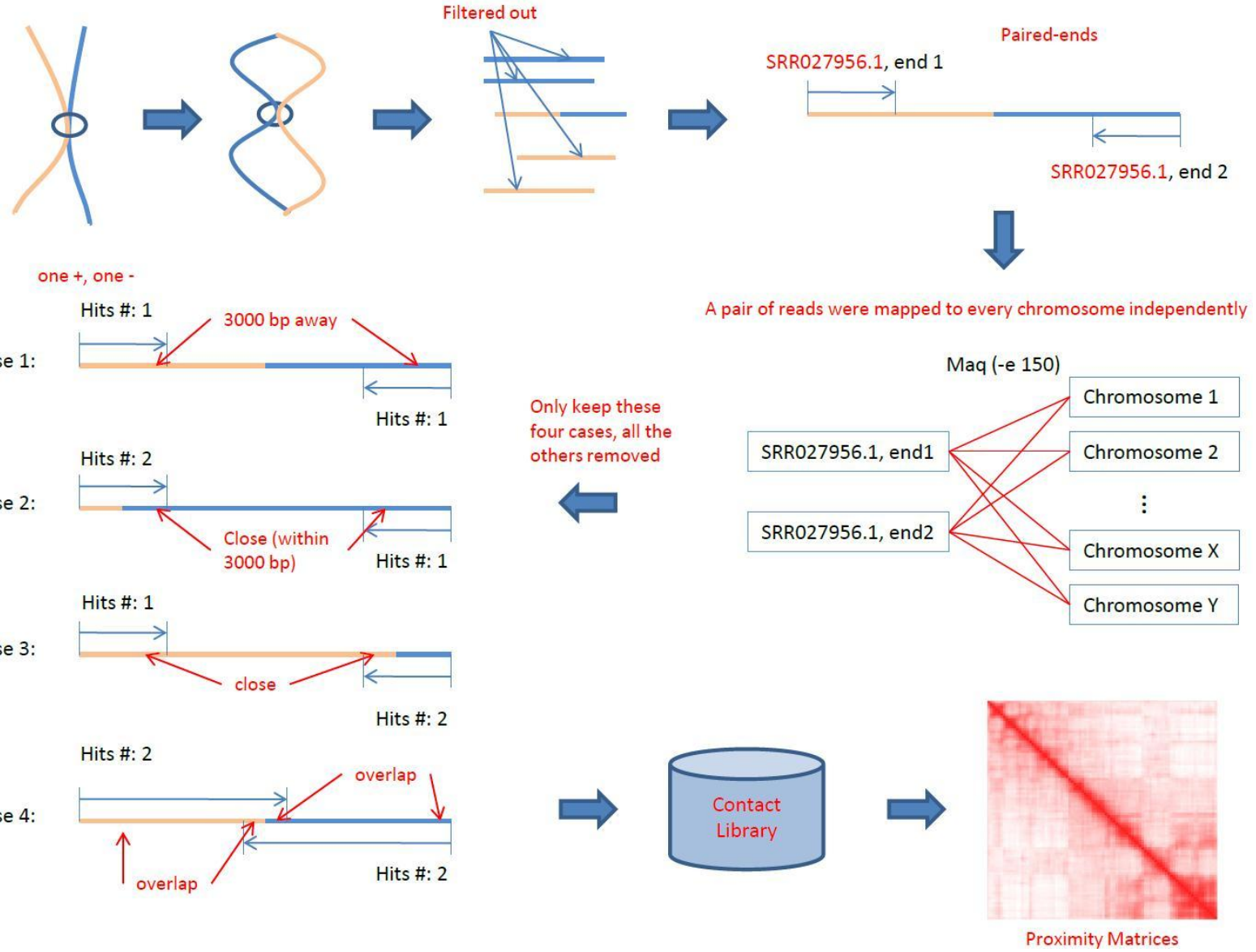
Chromosome Conformation Capturing (Hi-C)



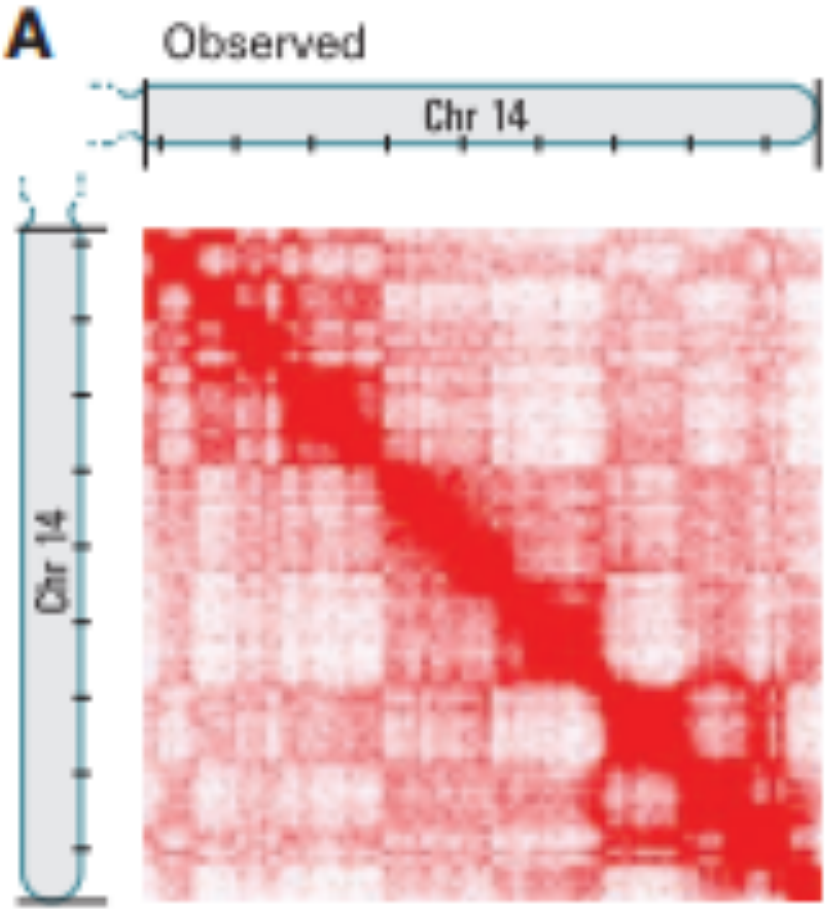
Hi-C Protocol

- Cells are crosslinked with formaldehyde
- DNA is digested with a restriction enzyme and ligated, resulting enriched with cross-linked elements with a biotin marked at the junction
- Shearing DNA and selecting biotin-containing fragments to create a Hi-C library
- Sequence the library to create a catalog of interacting fragments

Hi-C Data Analysis Pipeline

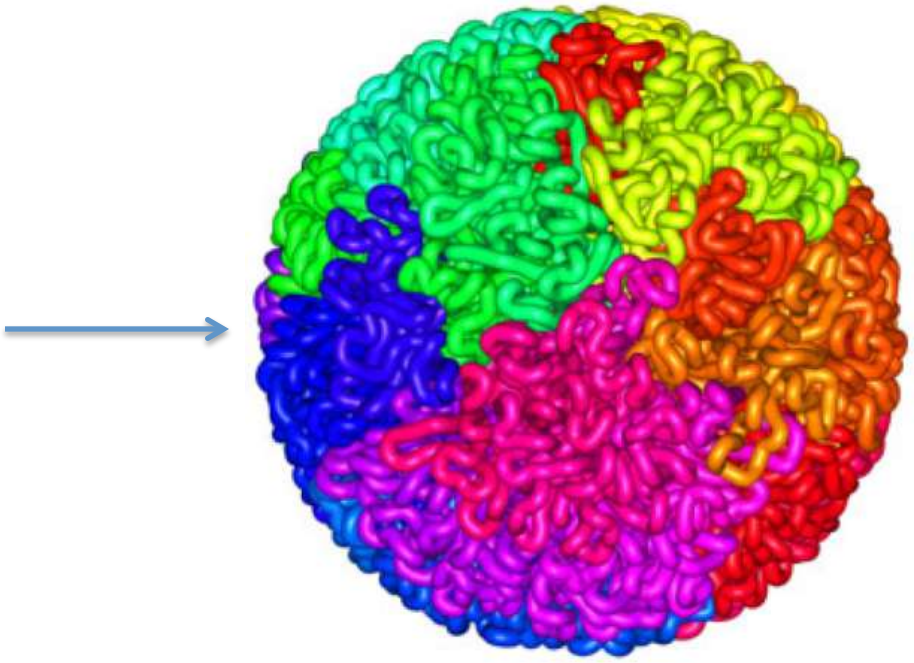


2D Chromosome Contact Map



Chromosome Conformation Capturing

Construct 3D Shape of Genome



Images.google.com

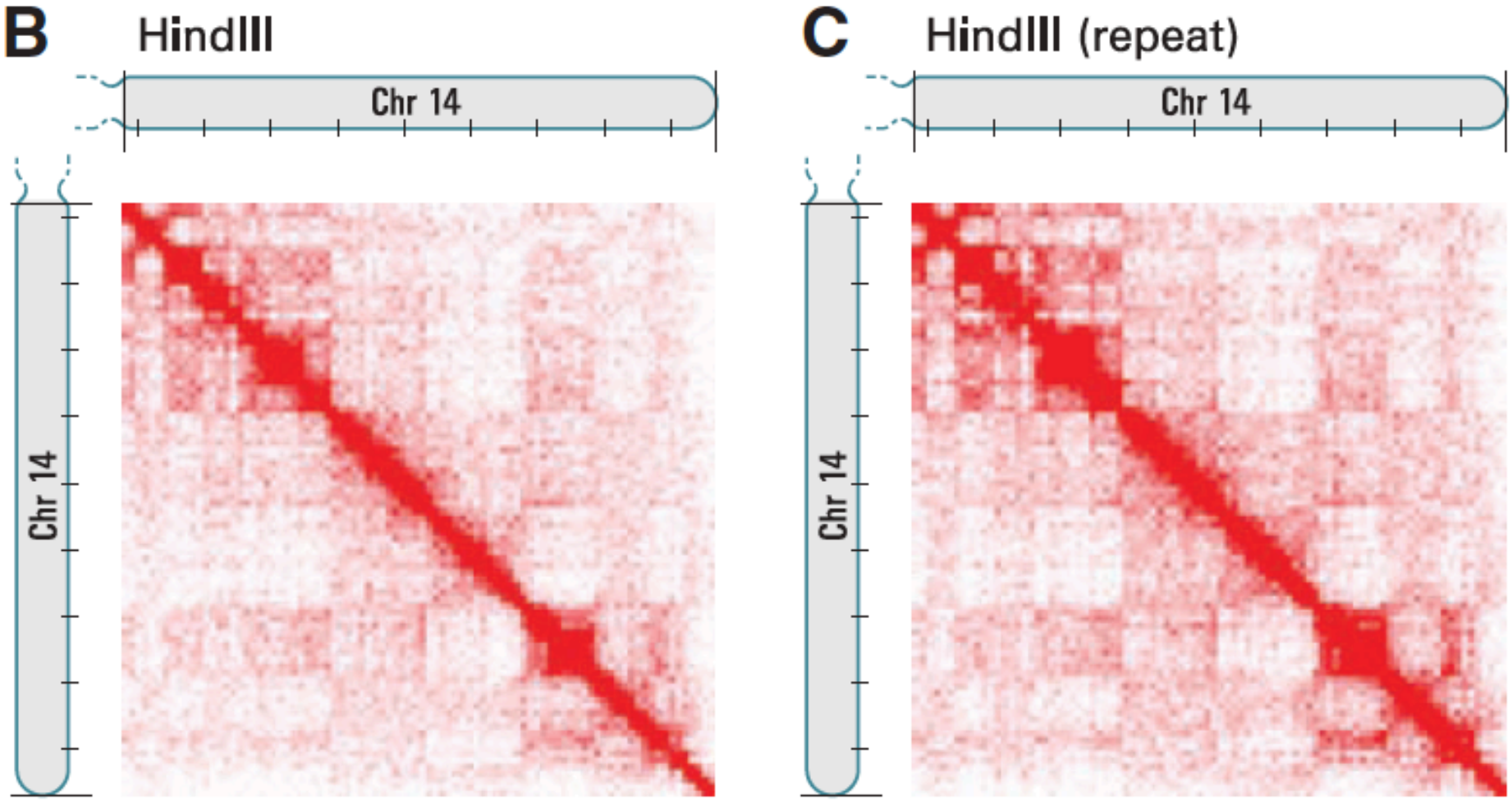
Data Set I

- A normal human lymphoblastoid cell line (B-cell)
- 8.4 million read pairs uniquely mapped to the human genome reference sequence
- 6.7 million corresponded to long-range contacts between segments > 20 kb apart

Liberman-Aiden et al., 2009

Genome Wide Contact Map

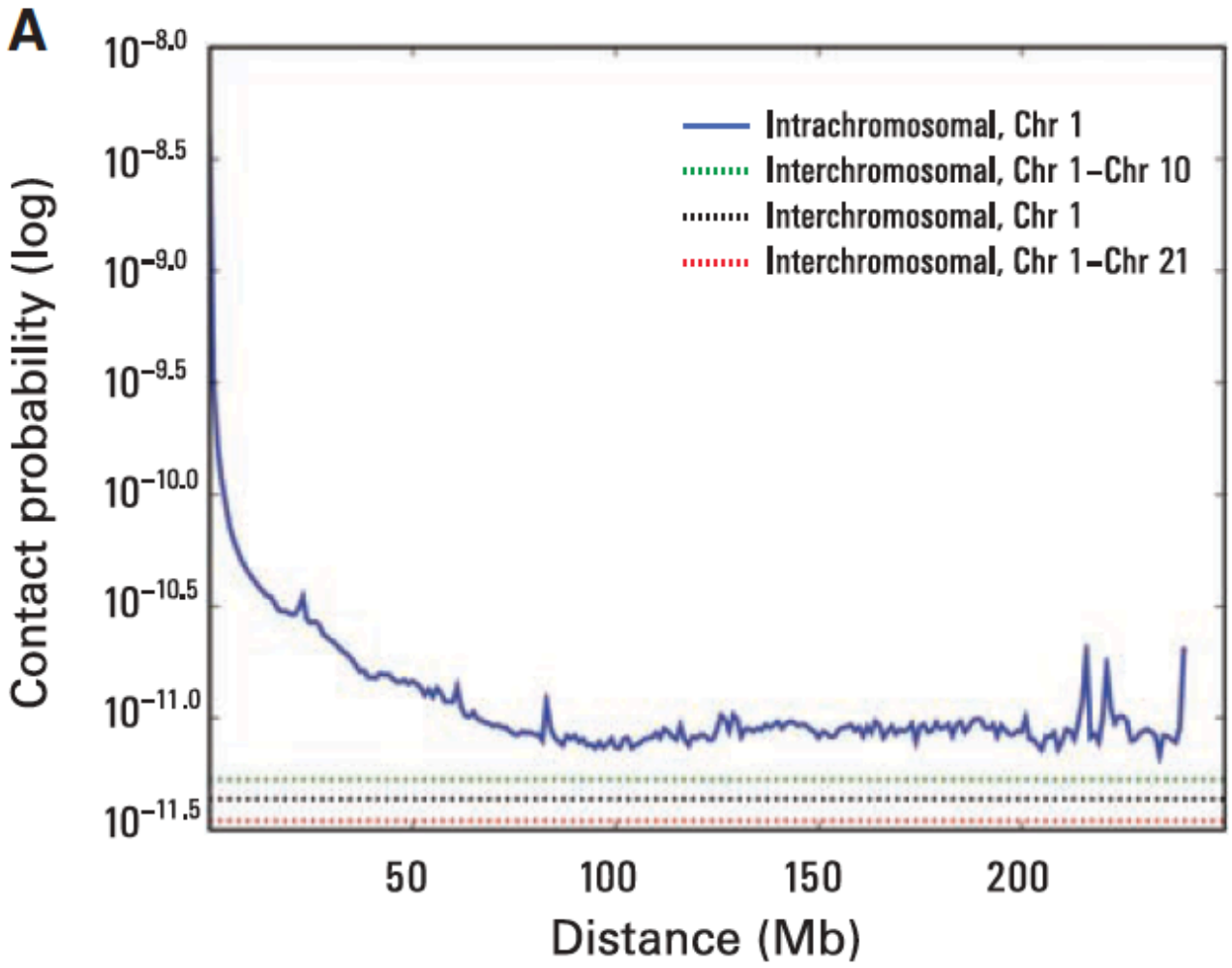
- Divide genome into 1-Mb regions (loci)
- M_{ij}: number of contacts between loci i and j
- The matrix reflects an ensemble average of the interactions in the original sample of cells
- Represented as a heat map



(B) Hi-C produces a genome-wide contact matrix. The submatrix shown here corresponds to intrachromosomal interactions on chromosome 14. (Chromosome 14 is acrocentric; the short arm is not shown.) Each pixel represents all interactions between a 1-Mb locus and another 1-Mb locus; intensity corresponds to the total number of reads (0 to 50). Tick marks appear every 10 Mb. (C) We compared the original experiment with results from a biological repeat using the same restriction enzyme [(C), range from 0 to 50 reads]. Correlation is 0.99.

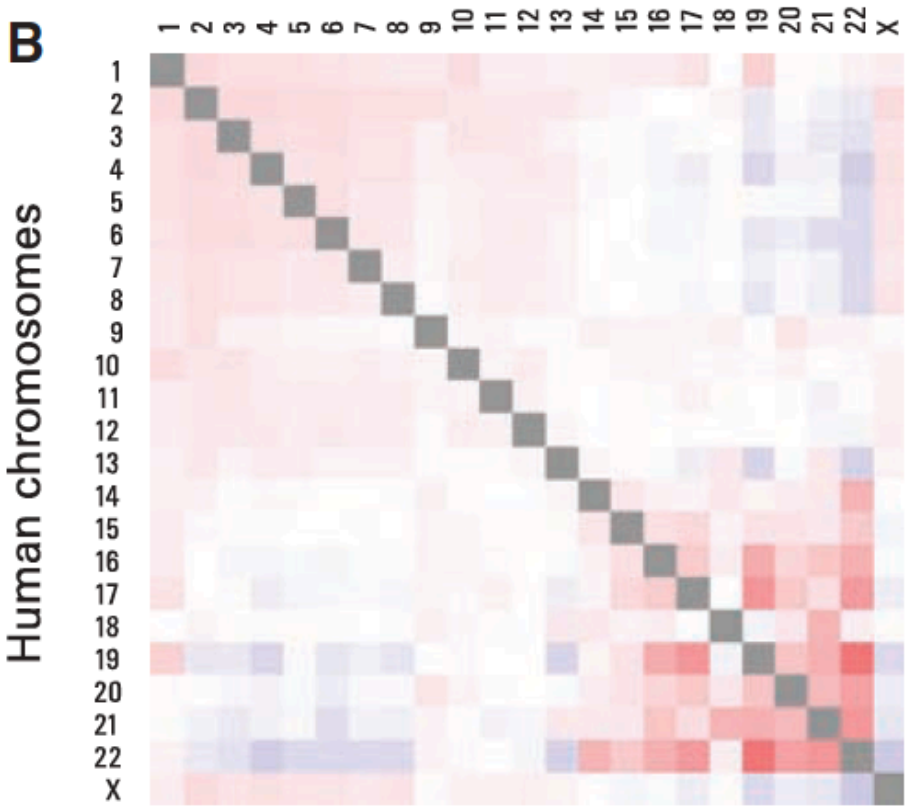
Relation between Euclidean Distance and Genomic Distance

- Average intrachromosomal contact probability $I_n(s)$ for pairs of loci separated by a genomic distance s on chromosome n.
- $I_n(s)$ decreases monotonically on every chromosome
- Even at distances > 200 Mb, $I_n(s)$ is always much greater than the average contact probability between different chromosomes



Implication: Chromosome territory

Probability of contact decreases as a function of genomic distance on chromosome 1, eventually reaching a plateau at ~90 Mb (blue). The level of interchromosomal contact (black dashes) differs for different pairs of chromosomes; loci on chromosome 1 are most likely to interact with loci on chromosome 10 (green dashes) and least likely to interact with loci on chromosome 21 (red dashes). Interchromosomal interactions are depleted relative to intrachromosomal interactions.



Observed/expected number of interchromosomal contacts between all pairs of chromosomes. Red indicates enrichment, and blue indicates depletion (range from 0.5 to 2). Small, gene-rich chromosomes tend to interact more with one another, suggesting that they cluster together in the nucleus.

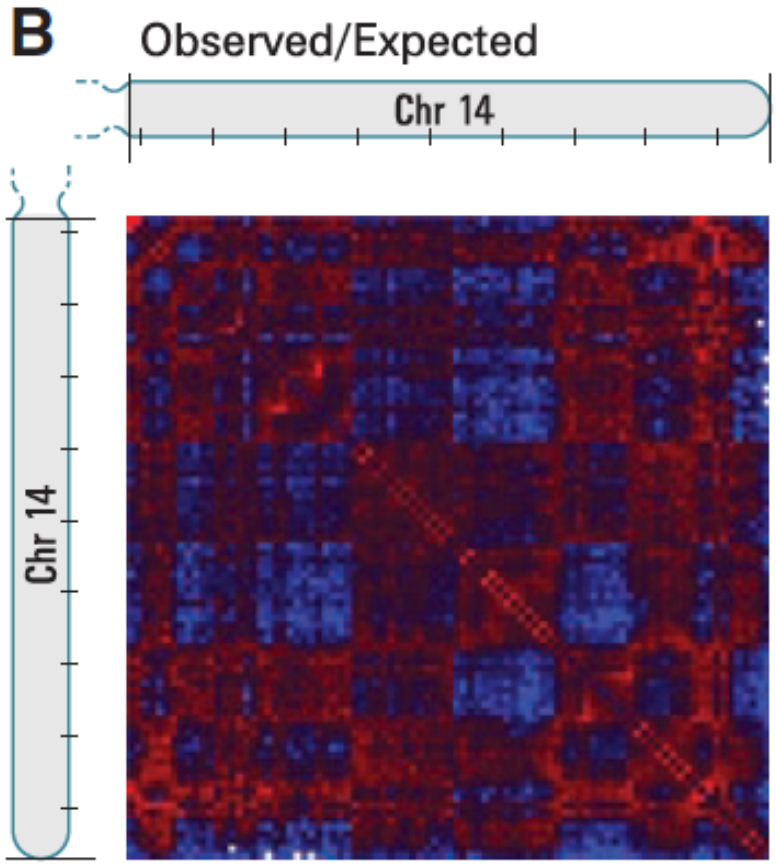
Human chromosomes

expected number of contacts between chromosome *i* and *j* was calculated by:

$$E_{i,j} = R_i \times R_j \times N_{INTER},$$

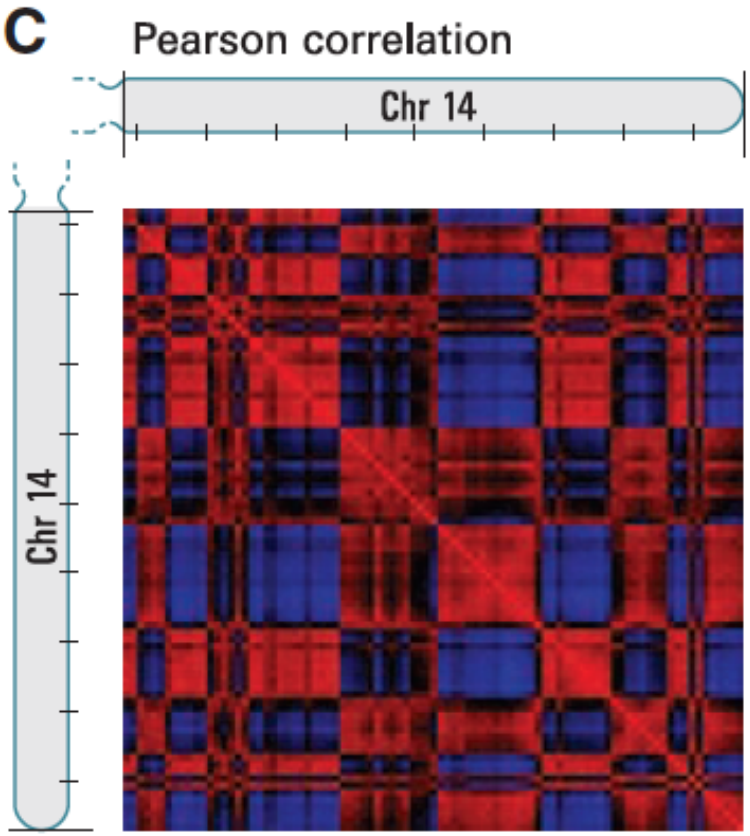
where R_i and R_j are the fractions of inter-chromosomal reads associated with *i* and *j*, respectively, and N_{INTER} is the total number of inter-chromosomal reads for a cell sample. The actual observed number of inter-chromosomal contacts between chromosomes *i* and *j* divided by the expected number $E_{i,j}$ indicates the enrichment or depletion of inter-chromosomal contacts between them.

Normalizing Contact Map by Expected Number of Contacts at Genomic Distance



M* by dividing each entry in the contact matrix by the genome-wide average contact probability for loci at that genomic distance. The normalized matrix shows many large blocks of enriched and depleted interactions, generating a plaid pattern

Pearson Correlation Map



If two loci (here 1-Mb regions) are nearby in space, we reasoned that they will share neighbors and have correlated interaction profiles.

This process dramatically sharpened the plaid pattern (Fig. 3C); 71% of the resulting matrix entries represent statistically significant correlations ($P \leq 0.05$).

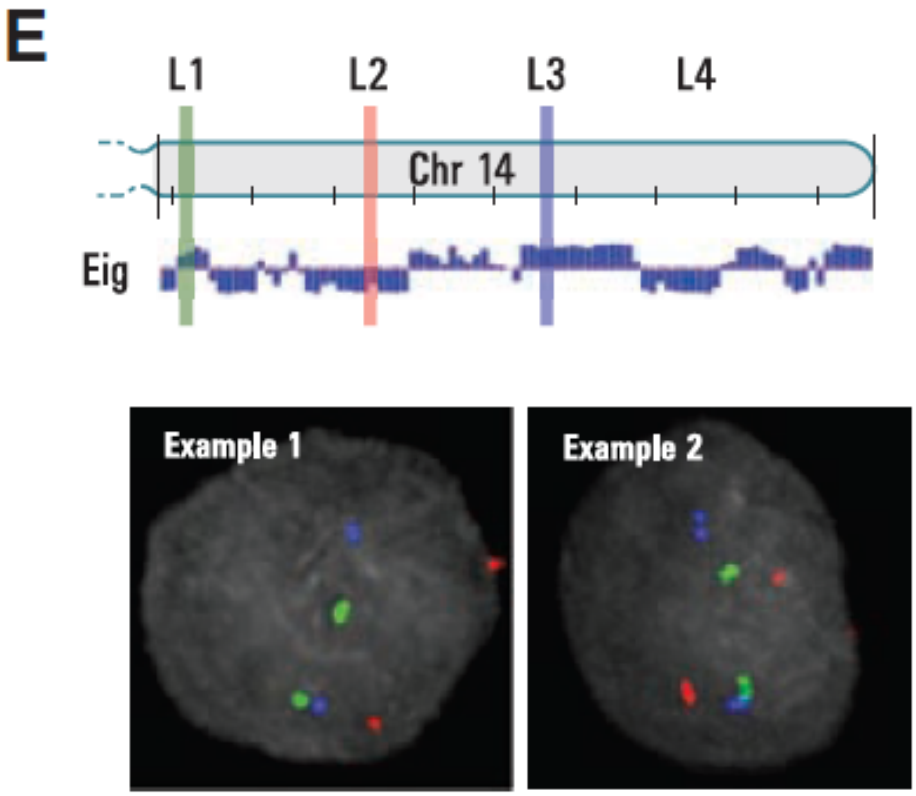
The **plaid pattern** suggests that each chromosome can be decomposed into two sets of loci (arbitrarily labeled A and B) such that contacts within each set are enriched and contacts between sets are depleted.

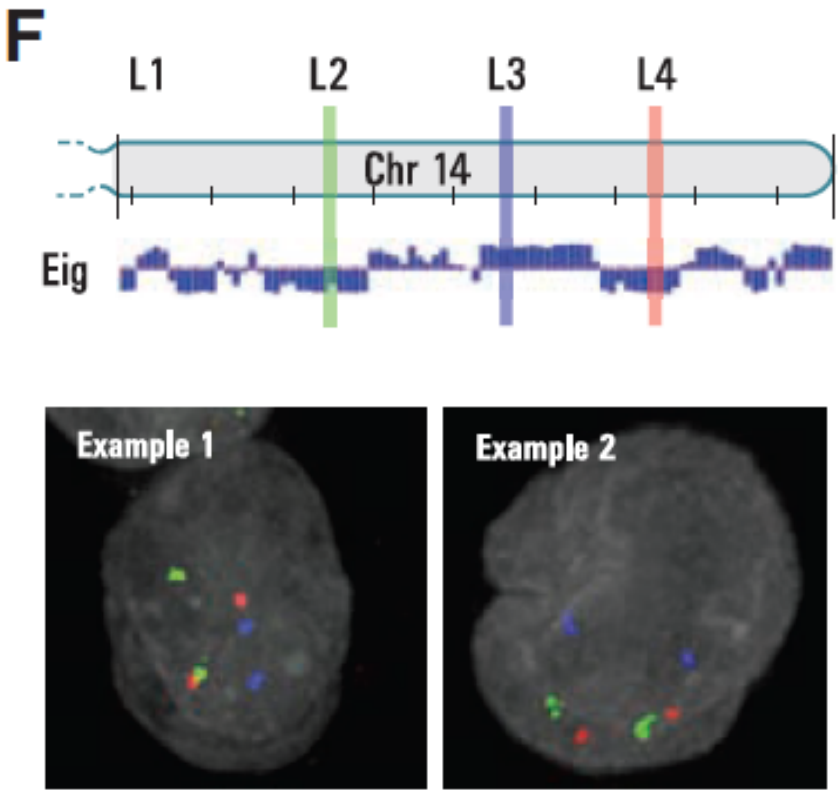
Principle Component Analysis

For all but two chromosomes, the first principal component (PC) clearly corresponded to the plaid pattern (positive values defining one set, negative values the other) (fig. S1). The entries of the PC vector reflected the sharp transitions from compartment to compartment observed within the plaid heatmaps.

Moreover, the plaid patterns within each chromosome were consistent across chromosomes: the labels (A and B) could be assigned on each chromosome so that sets on different chromosomes carrying the same label had correlated contact profiles, and those carrying different labels had anticorrelated contact profiles (Fig. 3D). These results imply that the entire genome can be partitioned into two spatial compartments such that greater interaction occurs within each compartment rather than across compartments. The Hi-C data imply that regions tend be closer in space if they belong to the same compartment (A versus B) than if they do not.

FISH Validation of Two Compartments



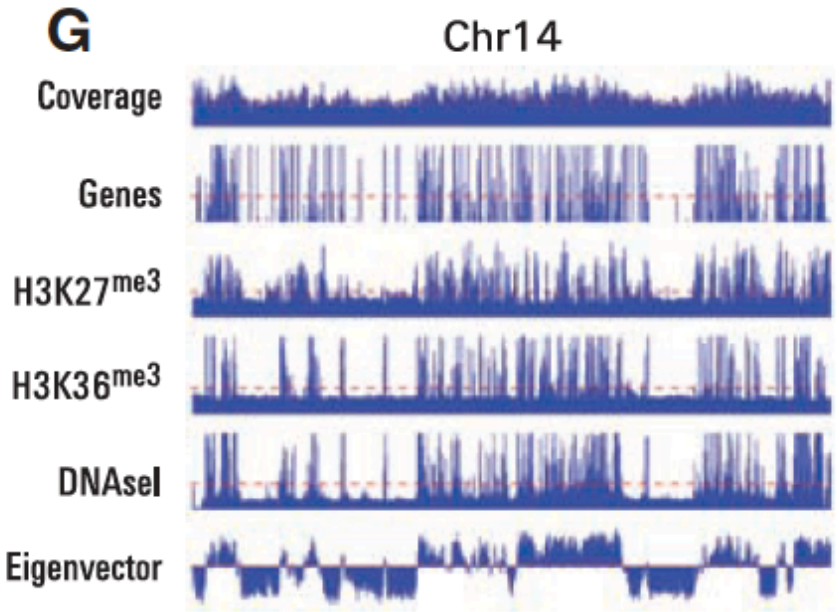


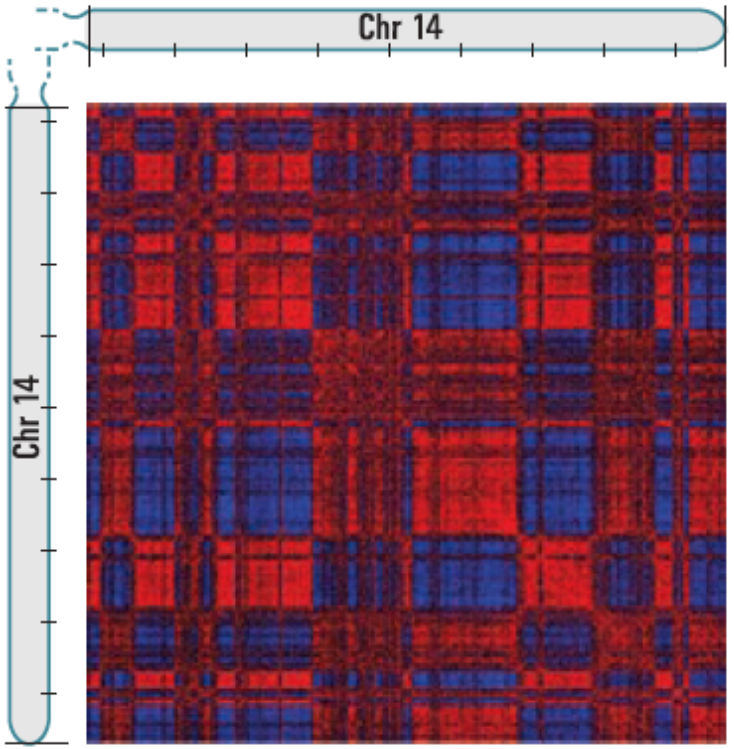
L1 – L3: Compartment A
L2 – L 4: Compartment B

Contacts and Physical Distance Measured by FISH

- More generally, a strong correlation was observed between the number of Hi-C reads m_{ij} and the 3D distance between locus i and locus j as measured by FISH [Spearman's r = –0.916, P = 0.00003 (fig. S3)], suggesting that Hi-C read count may serve as a proxy for distance.
- Upon close examination of the Hi-C data, we noted that pairs of loci in compartment B showed a consistently higher interaction frequency at a given genomic distance than pairs of loci in compartment A (fig. S4). This suggests that compartment B is more densely packed (15).
- The FISH data are consistent with this observation; loci in compartment B exhibited a stronger tendency for close spatial localization.

Open More Accessible, Gene Rich Compartment VS less Accessible



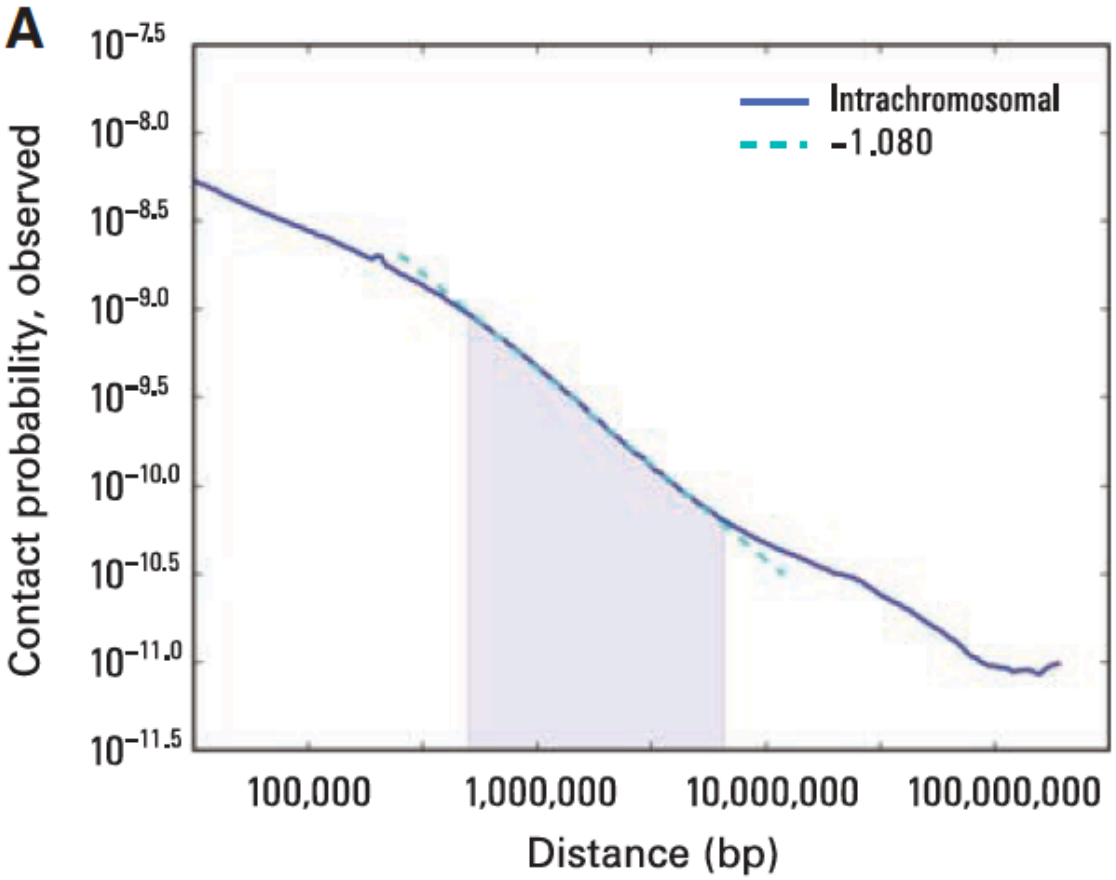


H3K27/H3K36: activating/repressive chromatin Marks

DNAseI: deoxyribonuclease I

(DNAseI) sensitivity, measure chromatin accessibility

Contact Probability VS Genomic Distance



Contact probability as a function of genomic distance averaged across the genome (blue) shows a power law scaling between 500 kb and 7 Mb (shaded region) with a slope of –1.08 (fit shown in cyan).

When plotted on log-log axes, I(s) exhibits a prominent power law scaling between ~500 kb and ~7 Mb, where contact probability scales as s^{-1} (Fig. 4A). This range corresponds to the known size of open and closed chromatin domains.

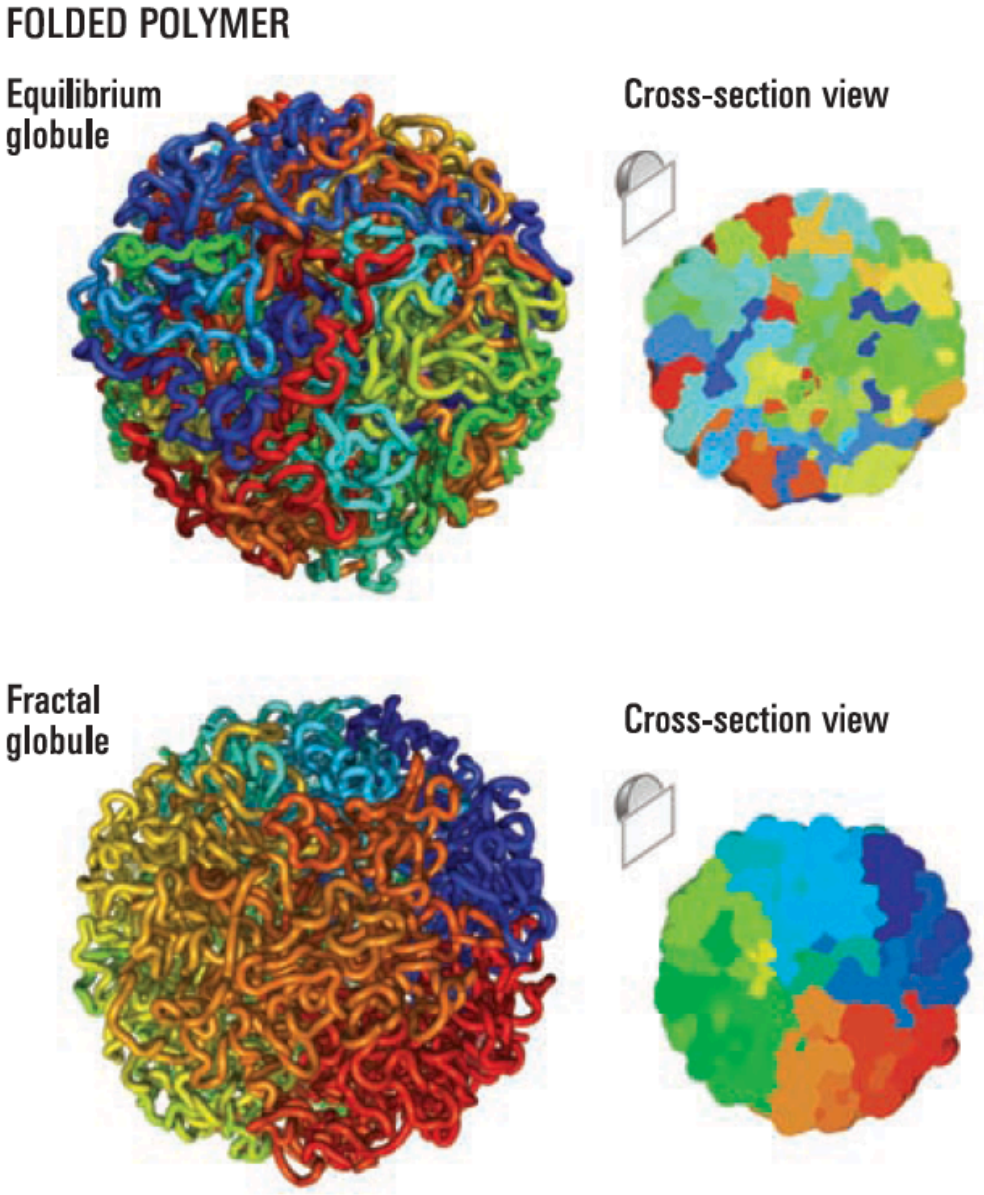
Genome 3D Model

- **Power-law** dependencies can arise from polymer like behavior.
- **Equilibrium globule**: a compact, densely knotted configuration originally used to describe a polymer in a poor solvent at equilibrium
- **Fractal Model**: This highly compact state is formed by an unentangled polymer when it crumples into a series of small globules in a “beads-on-a-string” configuration. These beads serve as monomers in subsequent rounds of spontaneous crumpling until only a single globule of globules-of-globules remains.

Fractal Model

- Fractal globules are an attractive structure for chromatin segments because they lack knots (31) and would facilitate unfolding and refolding, for example, during gene activation, gene repression, or the cell cycle.
- In a fractal globule, contiguous regions of the genome tend to form spatial sectors whose size corresponds to the length of the original region (Fig. 4C). In contrast, an equilibrium globule is highly knotted and lacks such sectors; instead, linear and spatial positions are largely decorrelated after, at most, a few megabases (Fig. 4C).

Comparison of Two Models

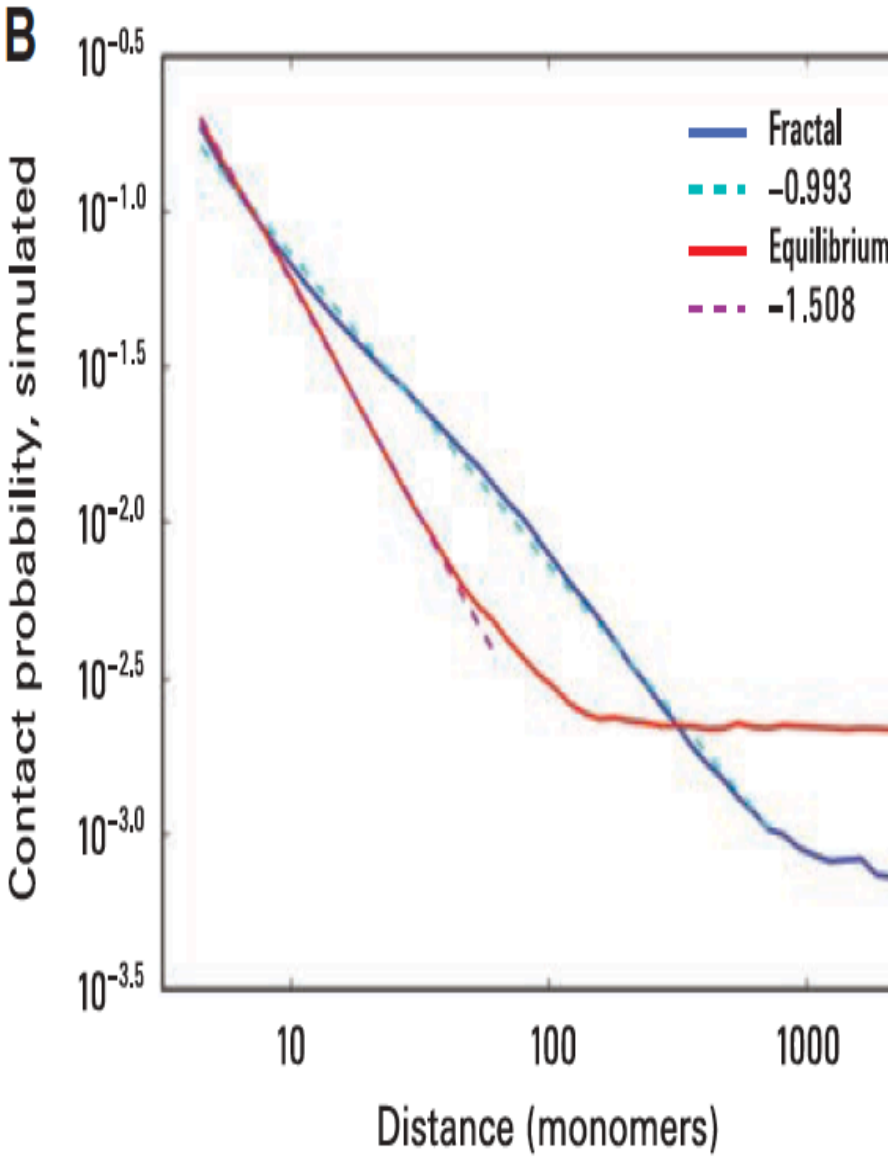


Consistency Checking

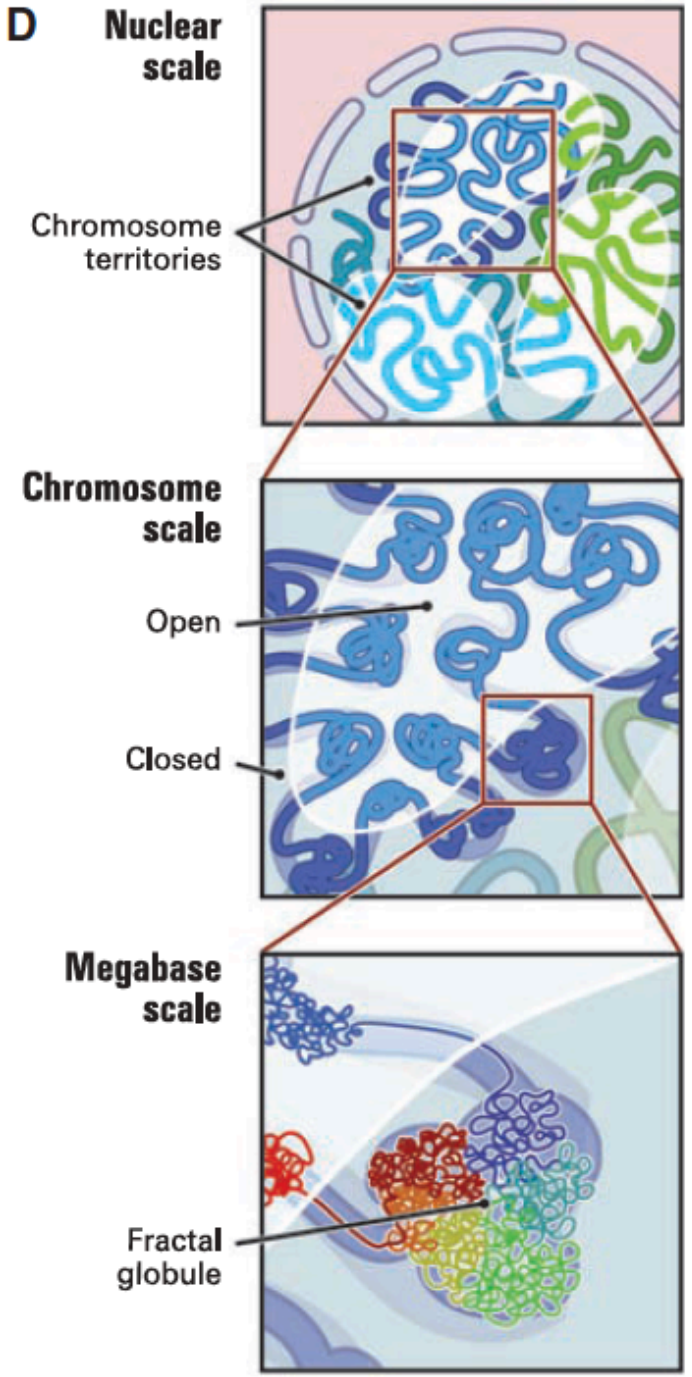
- The equilibrium globule and fractal globule models make very different predictions concerning the scaling of contact probability with genomic distance s. The equilibrium globule model predicts that contact probability will scale as $s^{-3/2}$, which we do not observe in our data. We analytically derived the contact probability for a fractal globule and found that it decays as s^{-1} (10); this corresponds closely with the prominent scaling we observed ($s^{-1.08}$).
- The equilibrium and fractal globule models also make differing predictions about the 3D distance between pairs of loci ($s^{1/2}$ for an equilibrium globule, s1/3 for a fractal globule). Although 3D distance is not directly measured by Hi-C, we note that a recent paper using 3D-FISH reported an $s^{1/3}$ scaling for genomic distances between 500 kb and 2 Mb

MCMC Simulation Validation

We used Monte Carlo simulations to construct ensembles of fractal globules and equilibrium globules (500 each). The properties of the ensembles matched the theoretically derived scalings for contact probability (for fractal globules, s^{-1}, and for equilibrium globules, $s^{-3/2}$) and 3D distance (for fractal globules $s^{1/3}$, for equilibrium globules $s^{1/2}$). These simulations also illustrated the lack of entanglements [measured by using the knot-theoretic Alexander polynomial (10, 32)] and the formation of spatial sectors within a fractal globule (Fig. 4B).



Multi-Scale Fractal Model



Hi-C Data Analysis II

Z. Wang, R. Cao, K. Taylor, A. Briley, C. Caldwell, J. Cheng. **The Properties of Genome Conformation and Spatial Gene Interaction and Regulation Networks of Normal and Malignant Human Cell Types**. PLoS ONE. 2013

Data Sets

- Primary human acute lymphoblastic leukemia (B-ALL) B-cell
- The MHH-CALL-4 B-ALL cell line (CALL4)
- The follicular lymphoma cell-line (RL)
- Sequenced by Illumina HiSeq 2000

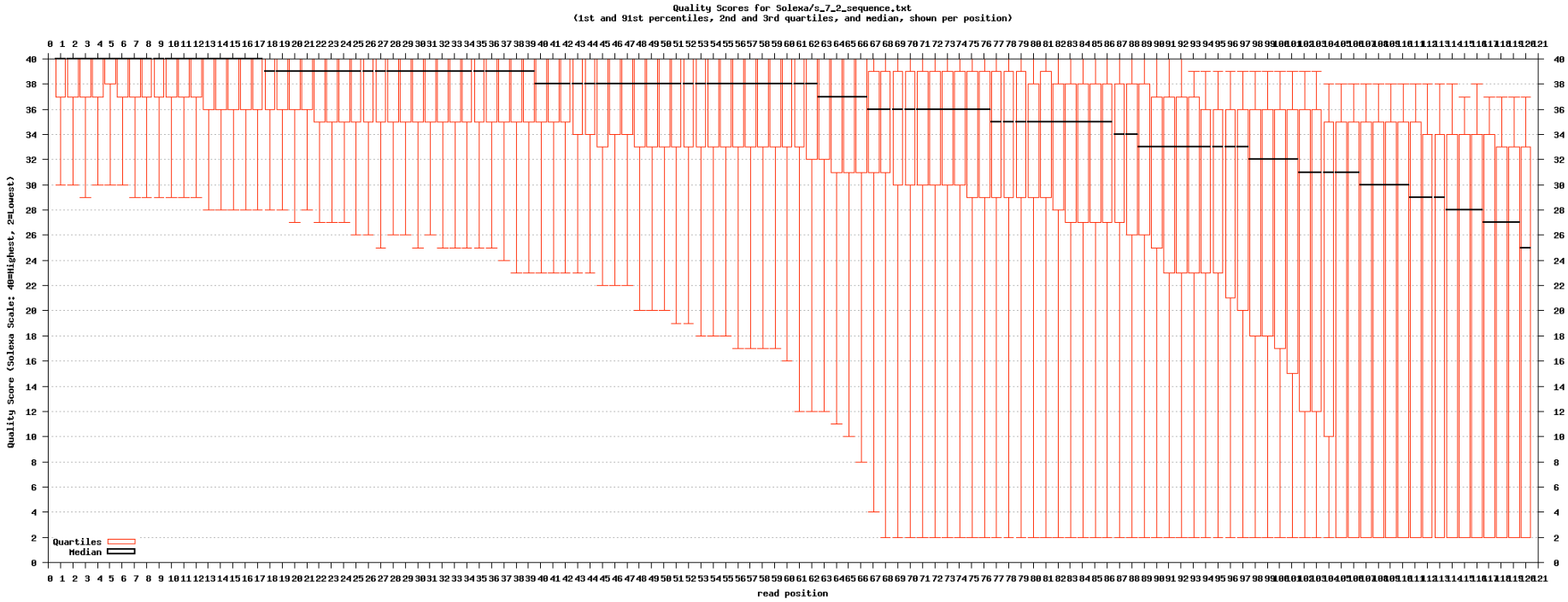
Number of Reads of the samples

Samples	Total number of reads	Utilized for analysis
Normal B cell	12,887,282	YES
RL1	60,272,006	NO
RL2	61,043,078	NO
RL3	65,579,872	NO
RL4	125,256,746	YES
Call4_1	62,741,712	NO
Call4_2	62,607,906	NO
Call4_3	133,542,778	YES
ALL B-Cell	77,888,742	YES

Read coverage

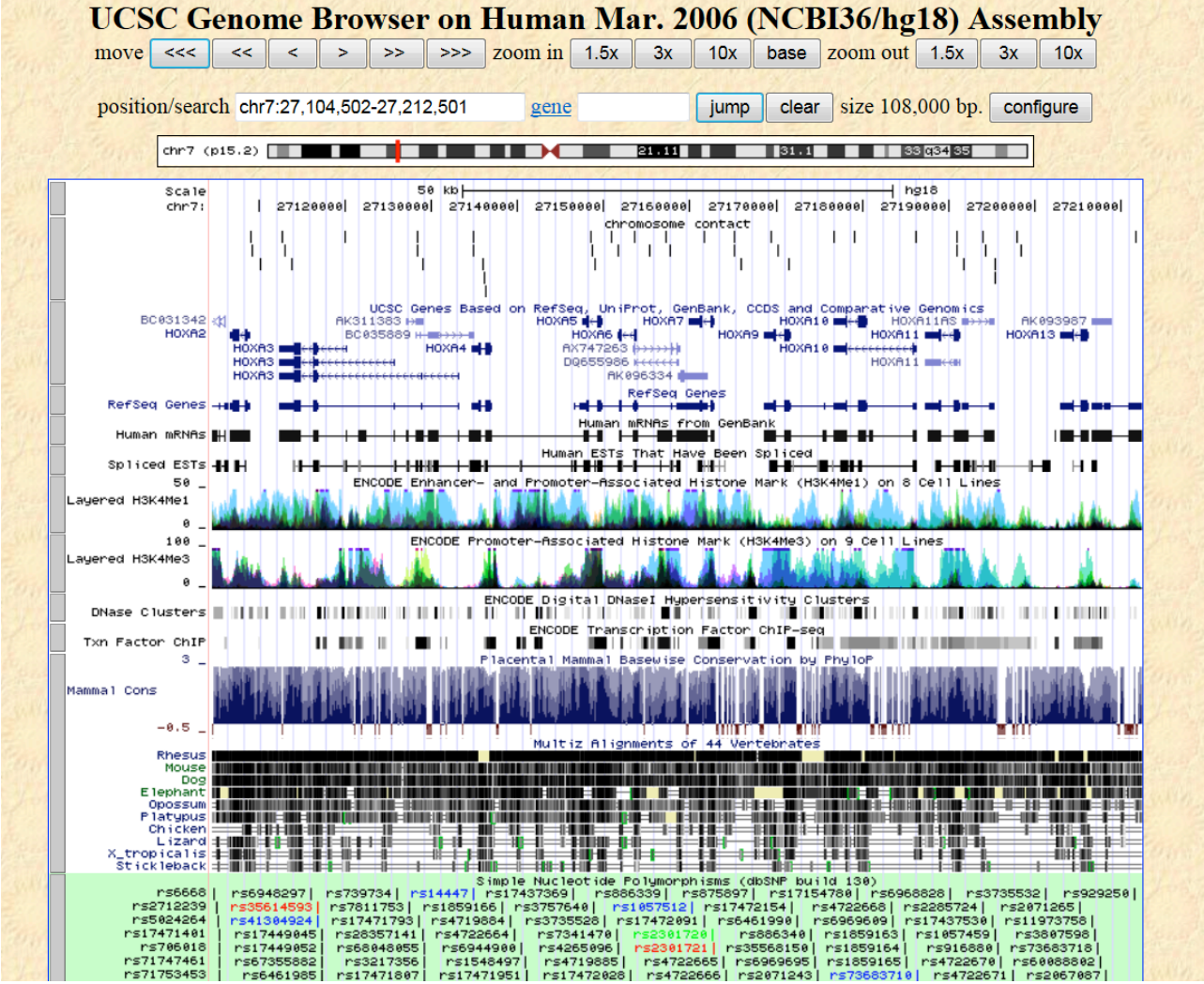
	Read coverage of gene region	Read coverage of non-gene region	Read length
Call4 cell line	2.81129121903121	2.35780895	100
RL cell line	1.47413416423764	1.14648699	100
Normal B-cell	0.186290512630489	0.1725446	76
ALL B-cell	1.788587532589	1.49037874	120

Read Quality

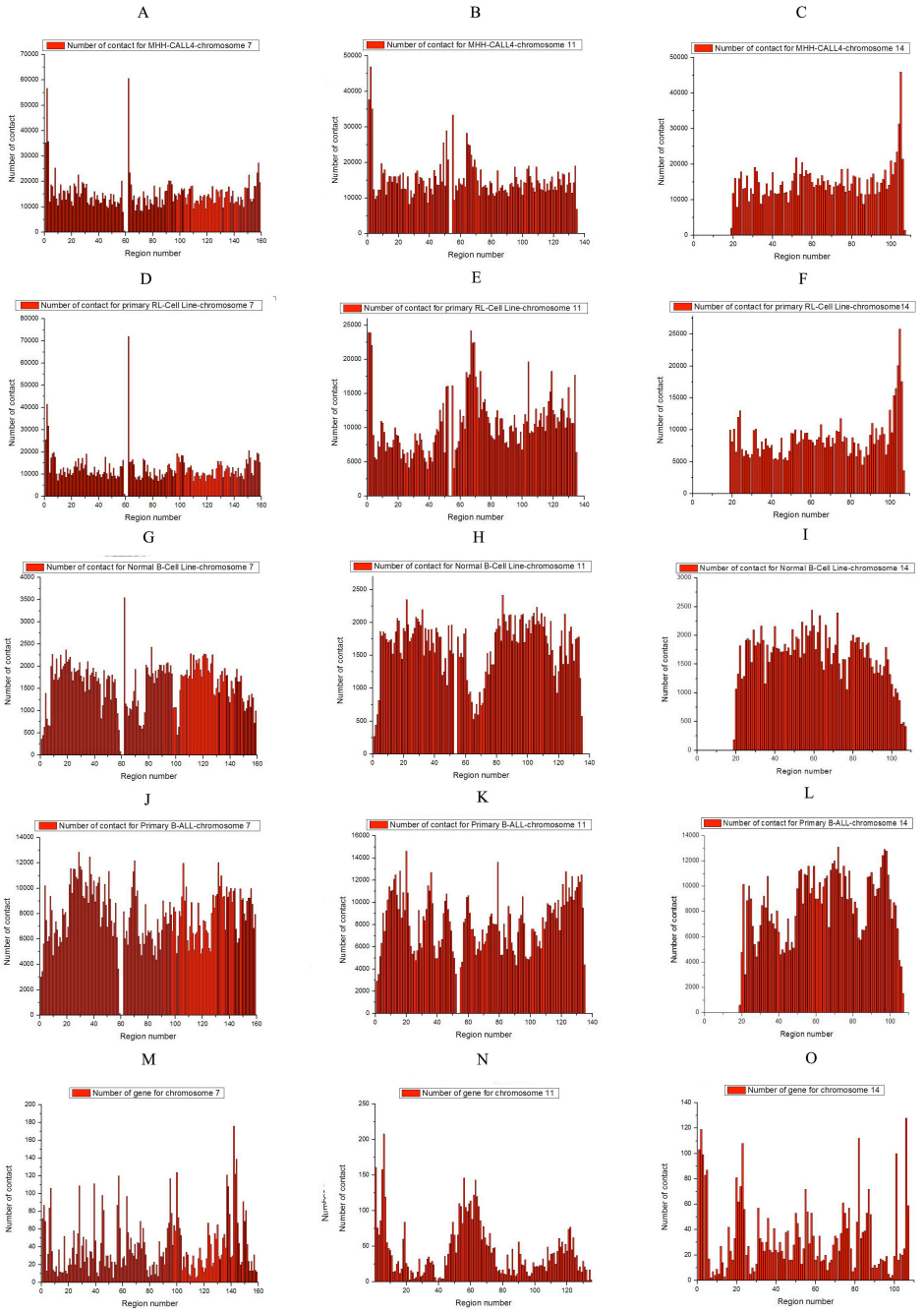


The sequencing quality score at a position is calculated as $Q_{solexa} = -10\log_{10}\frac{p}{1-p}$, where p is the probability of a sequencing error at the position. A score 30 means the probability of a sequencing error at the position is ~ 0.001. A score 20 or above may be considered acceptable.

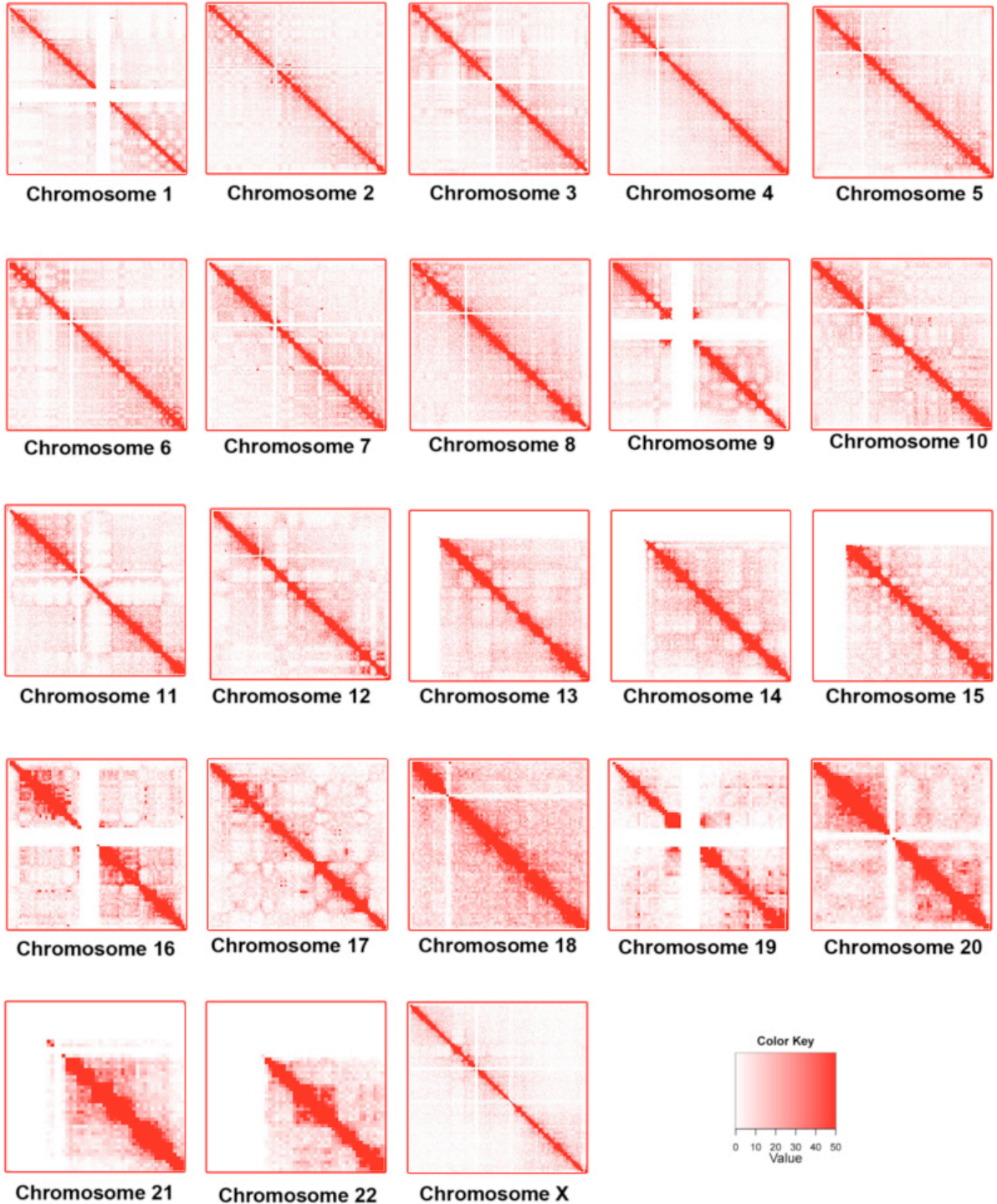
Mapping Reads to Human Genome



Plots of Contact Numbers against Regions of Chromosomes 7, 11, 14 (CALL4, RL, normal B-Cell, Primary B-ALL)



Un-normalized intra-chromosomal heat maps for primary ALL B-Cell



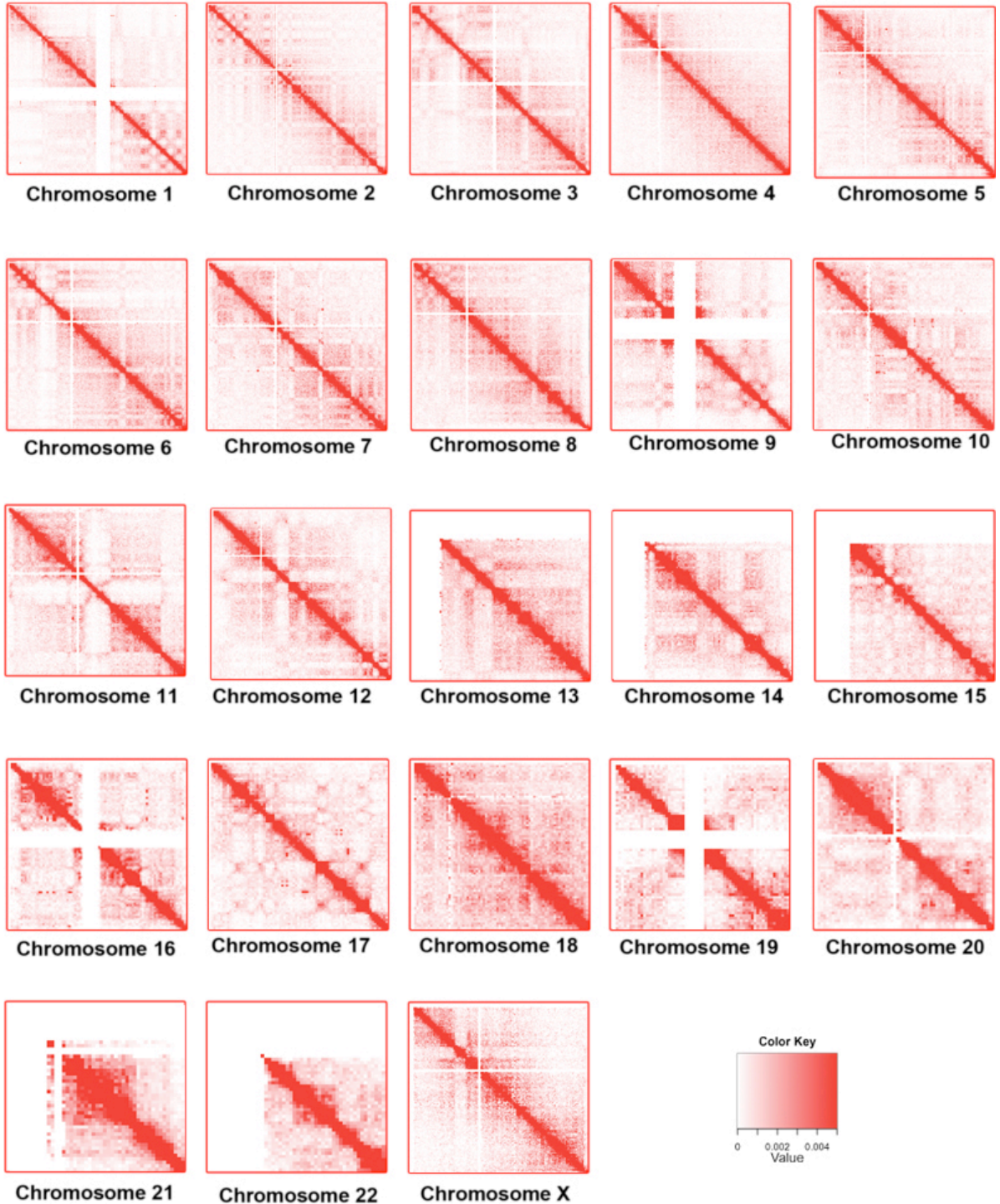
maps for primary ALL B-Cell

Sequential Component Normalization:

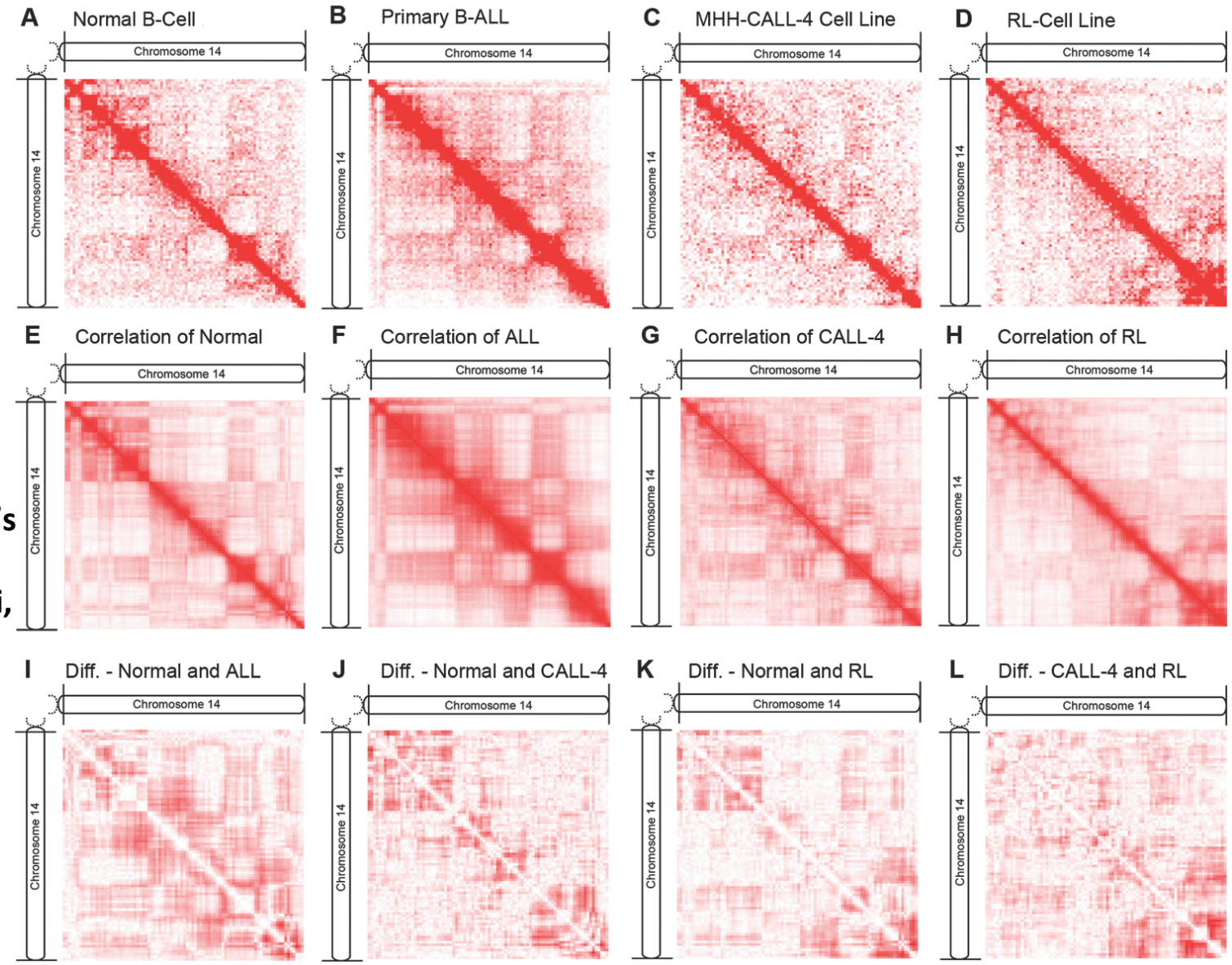
M[i,j] / |M[i]|

M[i,j] / |M[j]|

Repeat until symmetric



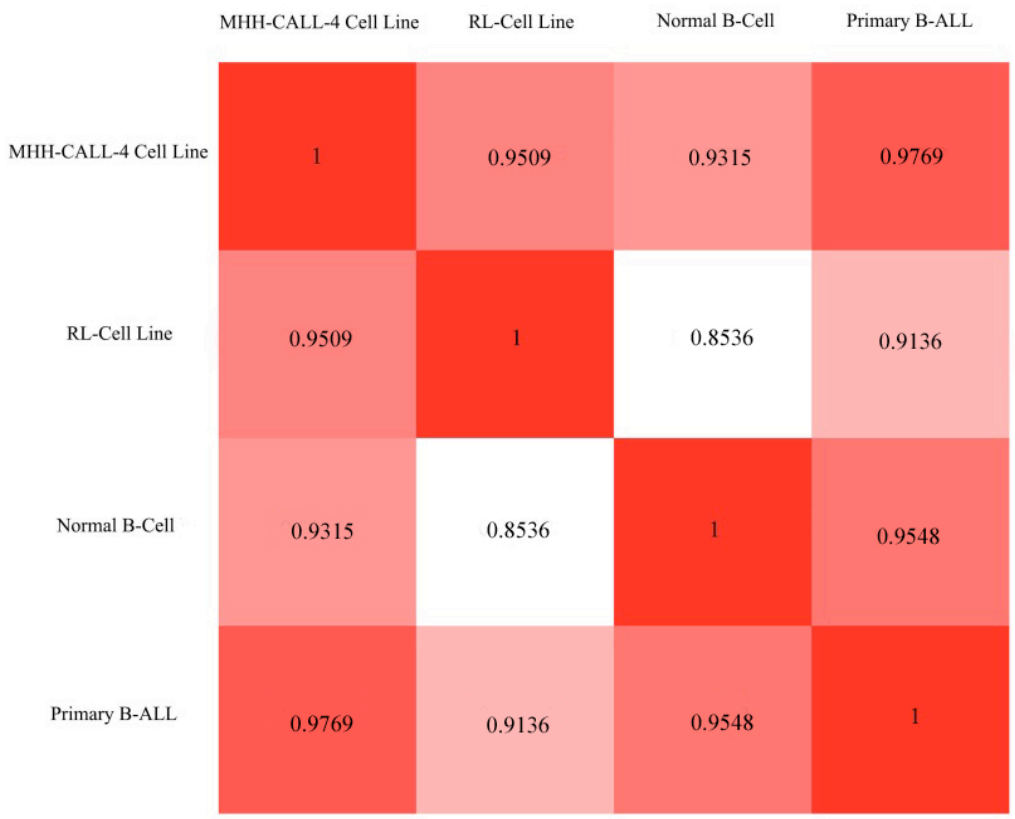
Intra-Chromosomal Contacts



Correlation Map
C[i,j] = Person's Correlation Between row i, column j

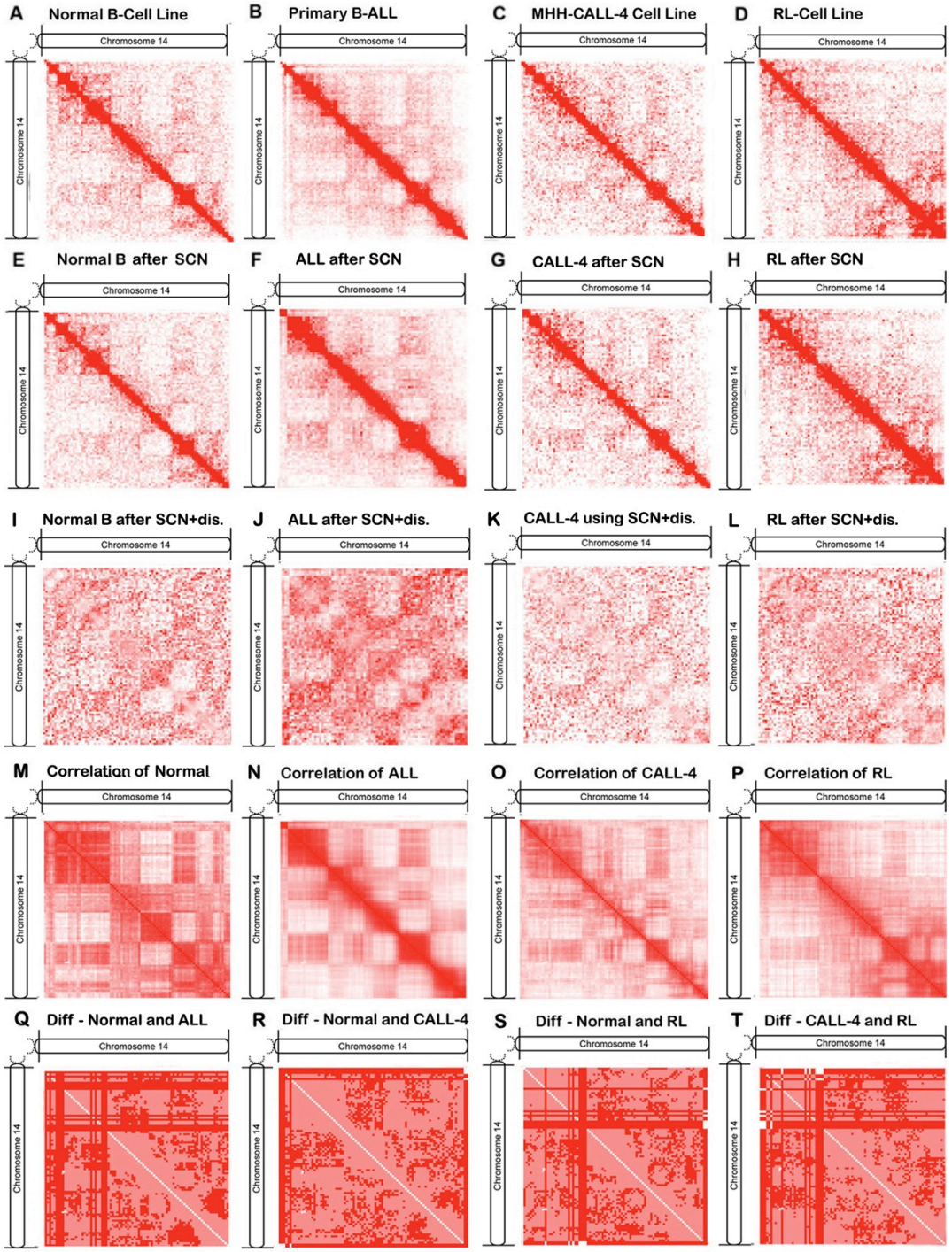
Plaid Pattern

Contact Correlation Between Cell Types

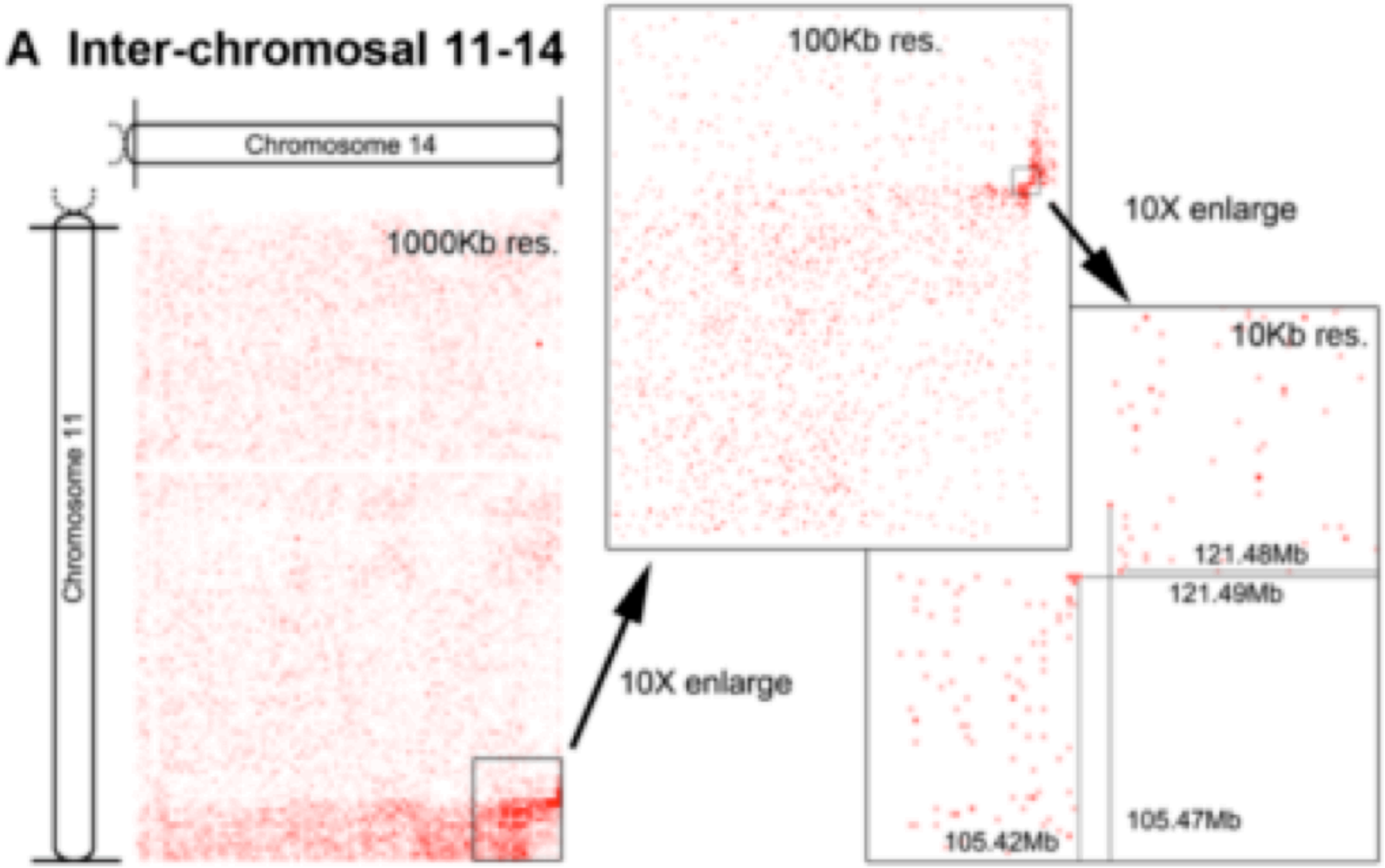


Correlate intra-contact numbers of 23 pairs of chromosomes

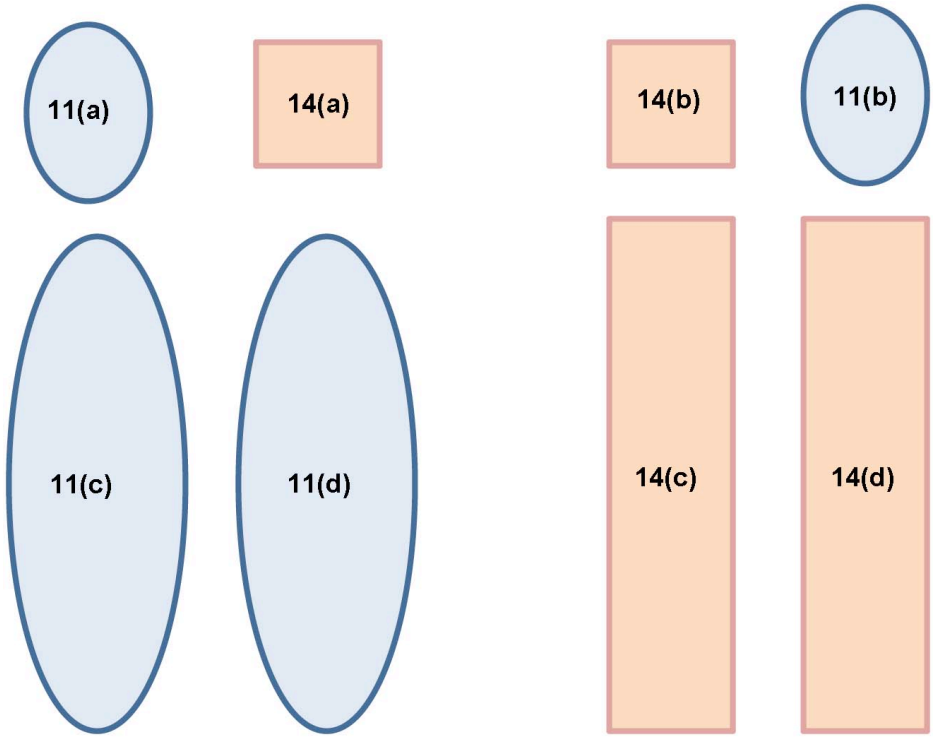
Comparison of Maps Using Genomic Distance and SCN Normalizations



Inter-Chromosome Contacts and Chromosome Translocation



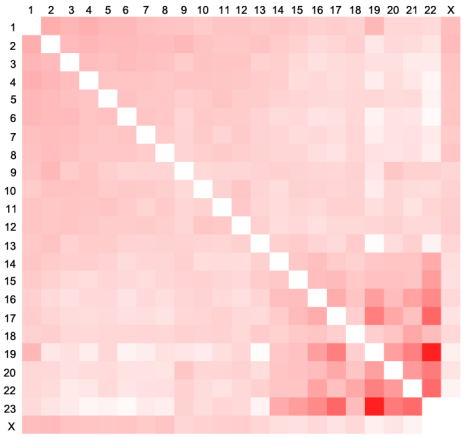
Cancer Causing Chromosome Translocation



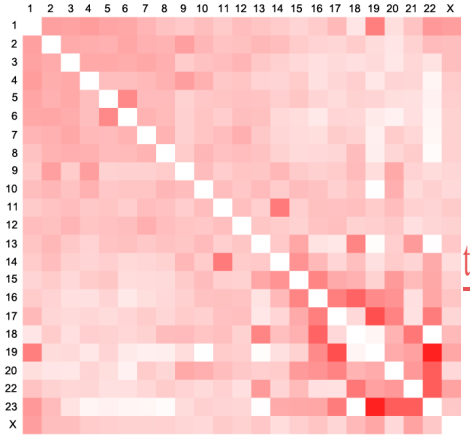
- Reconstruction of translocated chromosomes
- Two cancer related genes

Comparison of Inter-Chromosome Contact Profiles of Different Cells

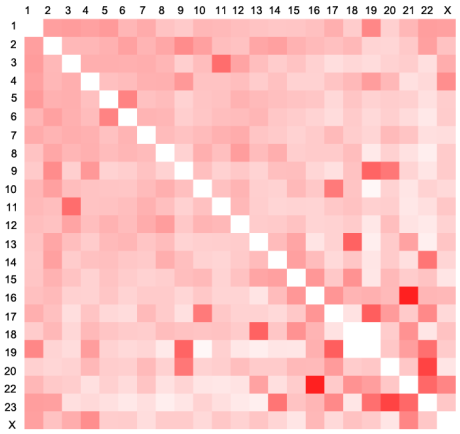
A Normal B-Cell Line



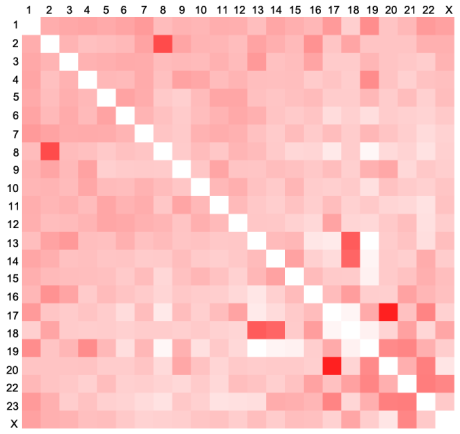
B Primary B-ALL



C MHH-CALL-4 Cell Line



D RL-Cell Line



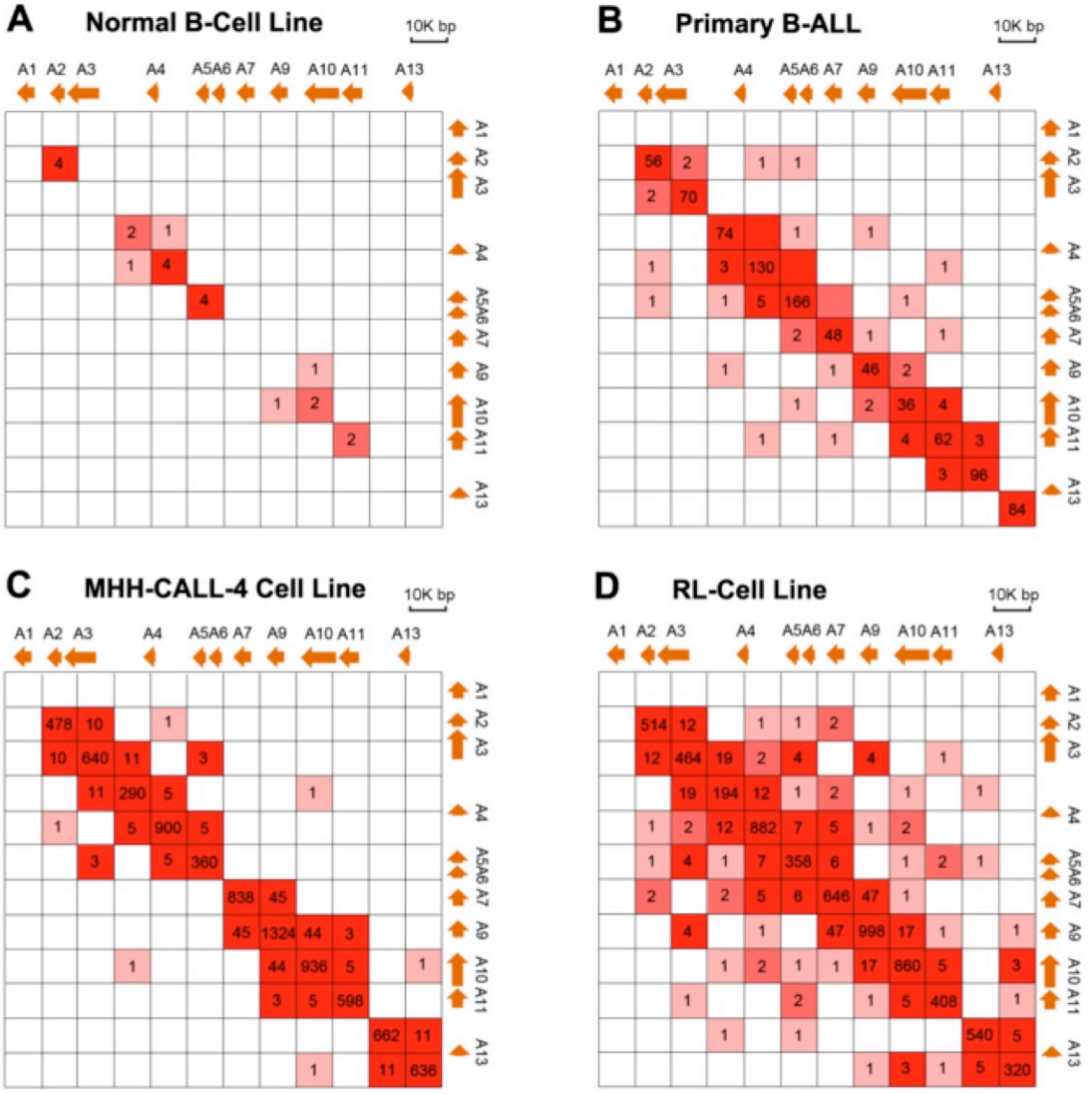
the expected number of contacts between chromosome *i* and *j* was calculated by

$$E_{i,j} = R_i \times R_j \times N_{INTER}$$

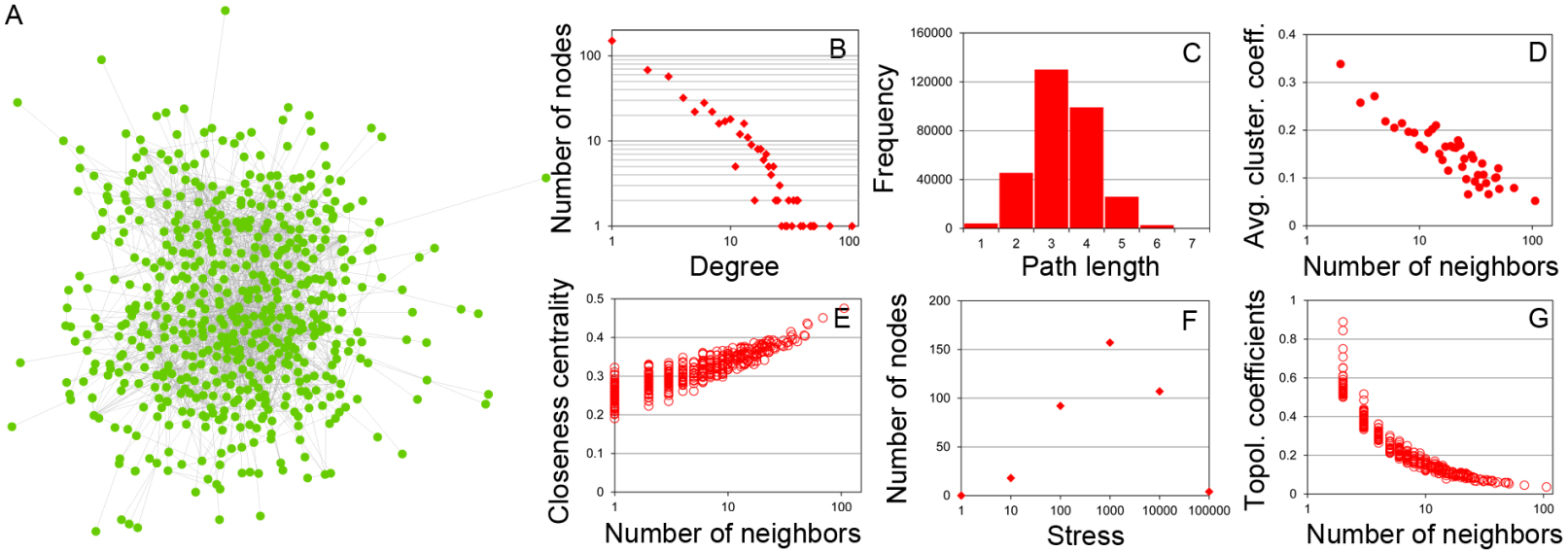
Gene-Gene Contact Map of HoxA Gene Cluster:

HoxA gene region (27,104,502 – 27,212,501) on chromosome 7,

Transcription factor controlling embrynoic development



Intra-Chromosome Gene-Gene Interaction Network of Chr. 14 of Call4 Cell Line



Applications?

Spatial gene interaction

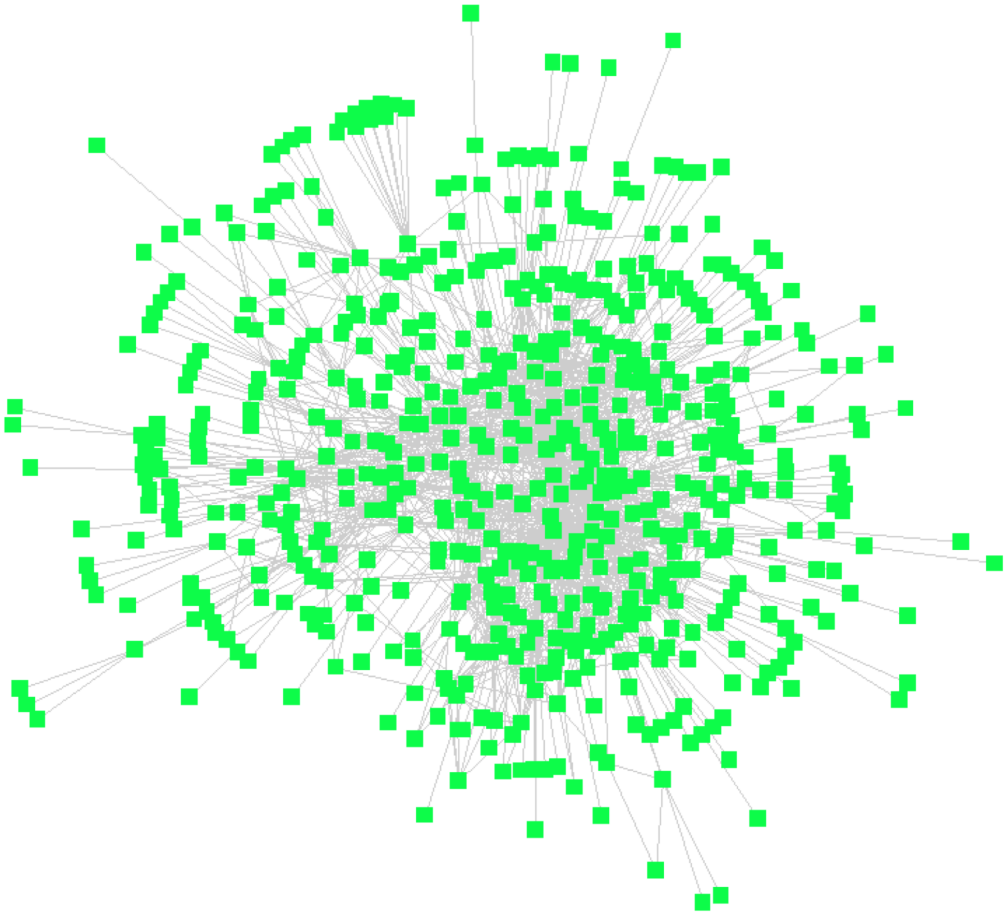
Gene function

Transcription efficiency

Scale free network: power law, short path, hub genes

Gene locations were determined by UCSC genome

Genome Wide TBS-TBS Interaction Network (Call4)

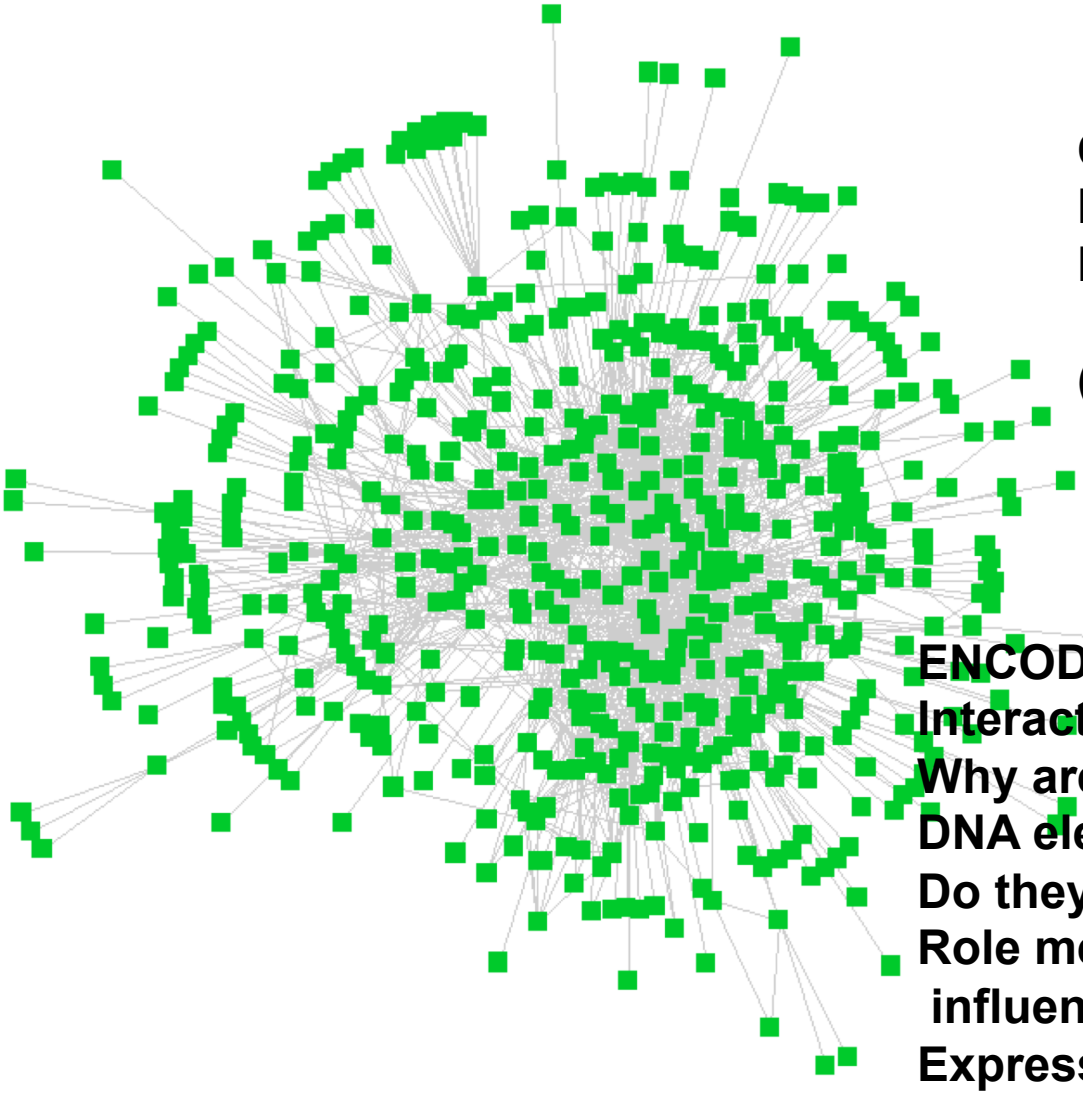


Question: why do TBSs form contact clusters and how?
Do they form control Module?

The definitions and coordinates of transcription factor binding sites were downloaded from Yale TFBS [31], which were identified by ChiP-seq experiments

Wang et al., submitted

Gene – TBS Interaction Network



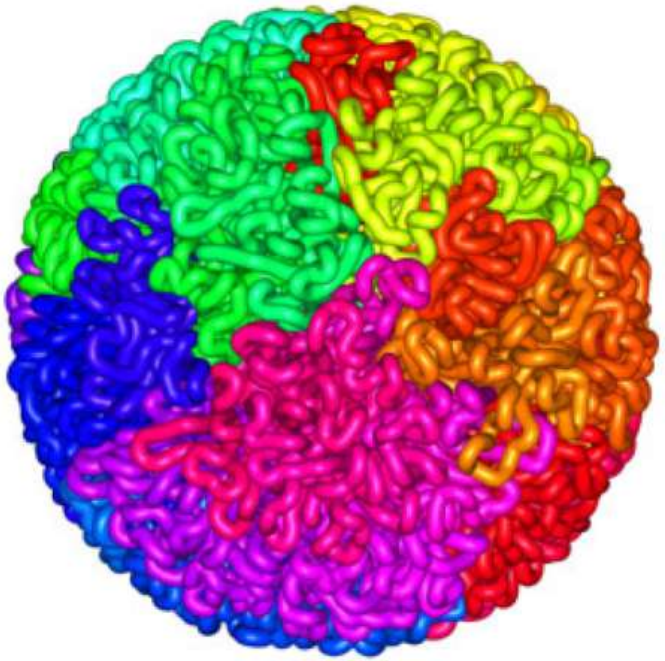
Question: can this explain long-range regulatory Relationships?

(promoter, enhancer, insulator)

ENCODE: Coding/Non-Coding Interactions are not random. Why are the non-coding DNA elements not junk? Do they play a structural Role mediated by interaction (insulator) influencing gene Expression at least?

Wang et al., 2013

3D Genome Structure Modeling (Key Hypothesis)



Images.google.com

Opportunities

- Chromosome contact data can be generated easily and cheaply
- Chromosome contact data is rather reliable
- A 3D model of a genome is very valuable in studying spatial regulation of gene expression and methylation

Challenges

- Genome and chromosome is very large (3 billion nucleotide of human genome)
- Genome structure is very dynamic
- No known experimental genome structure other than some point distance data generated by FISH
- Relationship between contact and distance is not deterministic
- Unknown quantities

Limitation of Existing Work

- **Small scale**: a gene locus or a small chromosome (10 – 50 M)
- **One Scale**
- **Rough approximation**: convert contacts to distance
- **Low accuracy**

A MCMC Approach

M. Rousseau, J. Fraser, M.A. Ferraiuolo, J. Dostie, M. Blanchette. ***Three-dimensional modeling of chromatin structure from interaction frequency data using Markov chain Monte Carlo sampling***. BMC Bioinformatics, 2011.

Long-range interactions between regulatory DNA elements such as enhancers, insulators and promoters play an important role in regulating transcription. As chromatin contacts have been found throughout the human genome and in different cell types, spatial transcriptional control is now viewed as a general mechanism of gene expression regulation.

MCMC5C

We formulate a probabilistic model linking 5C/Hi-C data to physical distances and describe a **Markov chain Monte Carlo (MCMC)** approach called MCMC5C to generate a representative sample from the posterior distribution over structures from IF data.

Structures produced from parallel MCMC runs on the same dataset demonstrate that our MCMC method mixes quickly and is able to sample from the posterior distribution of structures and find subclasses of structures.

MCMC5C

Structural properties (base looping, condensation, and local density) were defined and their distribution measured across the ensembles of structures generated. We applied these methods to a biological model of human myelomonocyte cellular differentiation and identified distinct chromatin conformation signatures (CCSs) corresponding to each of the cellular states.

MCMC5C

We also demonstrate the ability of our method to run on Hi-C data and produce a model of human chromosome 14 at 1Mb resolution that is consistent with previously observed structural properties as measured by 3D-FISH.

Existing Methods (Non MCMC)

- **5C3D (Fraser et al.)**: translates IF values into physical distance estimates and then uses a gradient descent approach to find the 3D conformations.

Existing Methods (Non MCMC)

- **Bau et al.** Interactions are modeled with springs whose equilibrium length depends on the observed IF values, subject to certain constraints based on the structure of the 30-nm fiber, optimized by Integrative Modeling Platform.

Existing Methods (Non MCMC)

- **Duan et al.,** convert interaction frequencies to Euclidean distances and then seek conformation minimizing the misfit, with addition of a set of clash avoidance constraints and a few prior known knowledge about the yeast genome organization. The constrained optimization problem is solved to find the best structure.
- **Tanizawa et al.** Similar to Duan et al.

Existing Methods (Non MCMC)

- **Common**: convert IF to distances, convert n*n constraints into n coordinates
- **Different**: conversion method, additional constraints, and optimization techniques.

Possible Drawbacks of Existing Methods

- Objective function (sum of square difference between predicted and derived distance) is debatable.
- Assume each IF is equally reliable.
- The absence of an underlying probabilistic model, preventing the calculation of confidence intervals on specific structural properties (e.g. distance between two genomic sites)

Probabilistic Model of Chromatin Conformation

- A chromosome is modeled as a continuous piece-wise linear curve in 3D.
- Theoretical interaction frequency between fragment i an d j, denoted IF(I, j), is inversely correlated with the distance between two fragments in 3D conformation: IF(i,j) = f(D_s(i,j)), where D_s(i,j) is the Euclidean distance between sites i and j and f is an appropriately chosen function.

$$f(D_S(i,j)) \propto 1/D_S(i,j)^{\alpha}$$

Observed IF and Theoretical IF

$$\widehat{IF}(i,j)|IF(i,j),\sigma(i,j)] =$$

$$N(\widehat{IF}(i,j); IF(i,j),\sigma(i,j)^2$$

re $N(x; \mu, \sigma^2)$ is the normal density functic

For 5C data, where σ can be estimated by IF-Calculator.

Observed IF and Theoretical IF (Hi-C)

$$\Pr[\hat{r}(i,j) \mid r(i,j)] = N(\hat{r}(i,j); r(i,j), r(i,j) + \kappa). \quad (2)$$

The role of κ, which we set to 10, is to avoid having small read counts being assigned too low a variance.

Prob(Structure | IF data)

The observed data $\widehat{IF}$ defines a posterior distribution over the set of possible conformations of the chromatin: $\Pr[\mathbf{S}|\widehat{IF}] = \Pr[\widehat{IF}|\mathbf{S}] \cdot \Pr[\mathbf{S}]/\Pr[\widehat{IF}]$. Since there are no constraints imposed on the structure space and the probability of the observed data $(\widehat{IF})$ is constant with respect to $\mathbf{S}$, we get $\Pr[\mathbf{S}|\widehat{IF}] = \zeta \cdot \Pr[\widehat{IF}|\mathbf{S}]$, for some constant ζ, and thus

$$\Pr[\mathbf{S}|\widehat{IF}] = \zeta \cdot \prod_{i,j} \Pr[\widehat{IF}(i,j)|IF(i,j) = f(D_{\mathbf{S}}(i,j), \sigma(i,j))].$$

Sampling Conformations from Posterior Distribution

- A random structure R_0 is initially chosen to seed the process (t=0), where each point is placed randomly in a cube of side length 10*avg(f(IF)).
- Repeat: The current structure R_t is randomly perturbed to obtain a new structure $R_{t'}$. If $Pr[R_{t'}|IF] > Pr[R_t|IF]$, the perturbation is obtained and we set $R_{t+1} = R_{t'}$. Otherwise, we set $R_{t+1} = Rt$.
- For values of t sufficiently large, $Pr[R_t=S] = Pr[S|IF]$.

Random Structure Perturbation

- Randomly choose one point S(i) along the structure and moving it by a vector v randomly choosing within a sphere of radius r (e.g. r = 0.25 nm) while keeping positions of other points unchanged – within a gene cluster
- The likelihood of the resulting structure is then quickly obtained from that of the old by updating the terms corresponding to the pairs of points involving i.

Assessing Mixing

- R_1, …, R_k of early iterations are highly dependent on R_0.
- Determine at what point m, the Markov process has mixed, i.e., R_m is independent of R_0
- After mixing, i.e. for k>= m, any sample R_k is representative of the target distribution. For d sufficiently large, R_k and R_{k+d} are independent.

Convergence Determination

- Run two independent chains R and R' in parallel, from independently chosen initial conformations R_0 and $R_{0'}$.
- Mixing is achieved if the samples $\{R_{k/2}, \ldots R_k\}$ and $\{R'_{k/2}, \ldots, R'_k\}$ cannot be distinguished from each other. Specifically, the average pairwise structural distances among R_k is compared to that between R_k and $R_{k'}$.
- After mixing is achieved, collect samples every d=k/20 iterations.

Clustering of Structure Ensembles

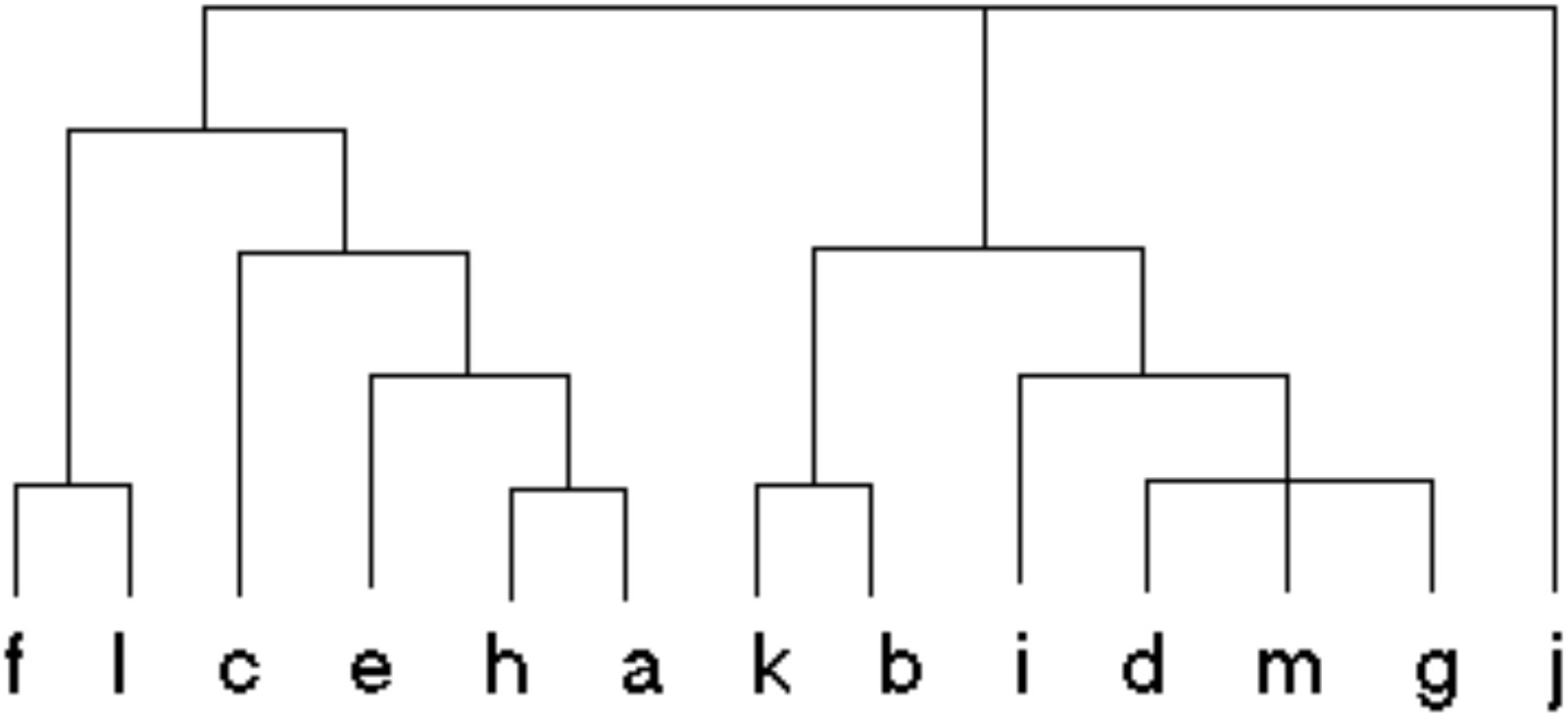
- Distance metric: N*N intra-structure distance matrix Ds.
- The distance (S,T) between two structures S and T is:

$$dist(\mathbf{S},\mathbf{T}) = \sqrt{\sum_{i,j} (D_{\mathbf{S}}(i,j) - D_{\mathbf{T}}(i,j))^2}$$

Structure Clustering

- Hierarchical clustering
- Visualization: tree dendrogram and a heatmap representation.
- Visual inspection is performed to determine the tree height cutoff and number of subfamilies
- Choose maximum likelihood structure from each cluster as representative and assigning it a weight proportional to the number of structures in its cluster.

Hierarchical Clustering



Measuring Structural Properties

- **Local density** at position i along the sequence is the number of DNA bases located within a sphere of radius r centered at position i.
- **Compaction**: the number of DNA bases located within the sphere and consecutive to position i.
- **Looping** counts the number of bases inside the sphere but outside the portion the sequence portion containing i.

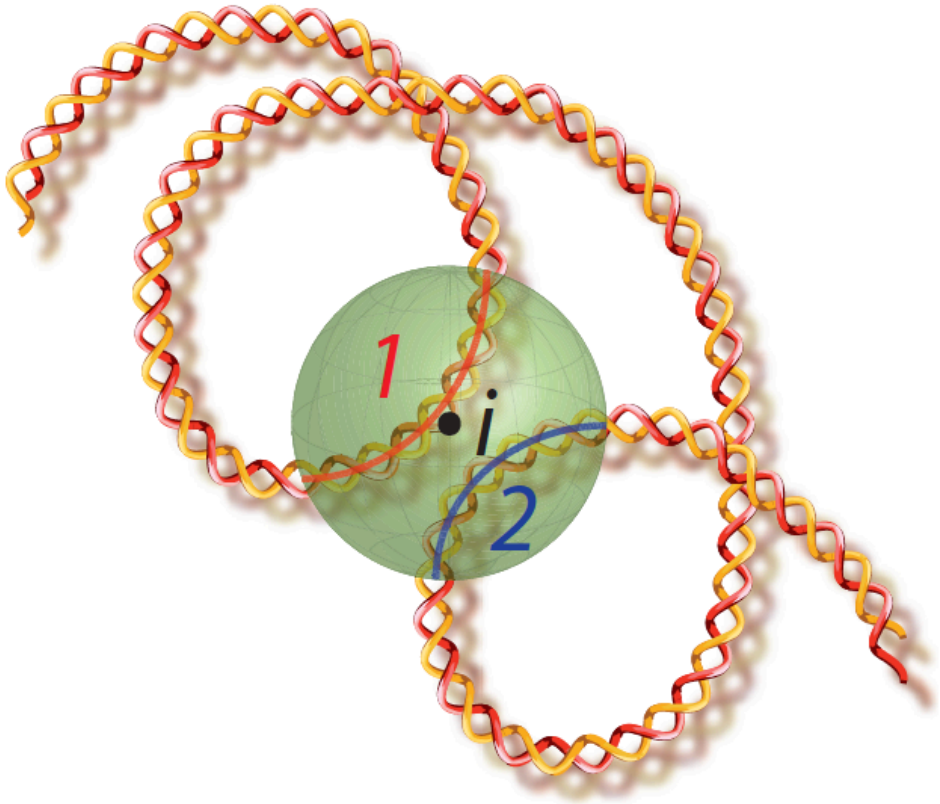


Figure 2 Structural Properties. Schematic diagram of Structural Properties. The shaded sphere with radius *r* is centered at base *i*. The nucleotides that lie within the sphere and delineate compartment 1 (nucleotides consecutive to base *i* before leaving sphere, indicated with a red arc) are counted as the base condensation measure and the nucleotides that lie within the sphere and delineate compartment 2 (nucleotides on sequence that has exited and re-entered sphere, indicated with a blue arc) are counted as the base looping measure. The total number of nucleotides contained within the sphere is counted as the base density measure.

5C Data Sets

- HoxA Gene Cluster
- In both datasets, the genomic, region analyzed spans **142** kb and contains **11** protein coding genes. The region contains **42** restriction sites, for the BglII restriction enzyme, which was used for the, experiment.
- MLL-ENL fusion cell line (HB-1119)

Convert IF to distance

- Duan et al., the resulting conversion approximately follows $d \propto 1/IF$. Mateos-Langerak et al . [50] also suggest a relationship of the form $d \propto 1 / IF^{a}$. Bau et al . [28] convert their IF via a linear transformation of the IF' s z-score. Tanizawa et al . [29] relate IF to physical distance by using a loess regression on a set of physical distances measured by 3D-FISH, but do not report the parameters of this regression.

Convert IF to distance

- The most accurate model ($d \propto C / IF^{a}$) is the one that is best able to predict unseen pairwise interaction frequencies. For each of a set of possible a leave-one-out cross-validation was performed. Find a to minimize

$$MSE(\alpha) = \frac{1}{n} \sum_{(i,j)} (D_{\mathcal{S}^{*}_{(i,j);\alpha}}(i,j)^{-\alpha} - \widehat{IF(i,j)})^{2}.$$

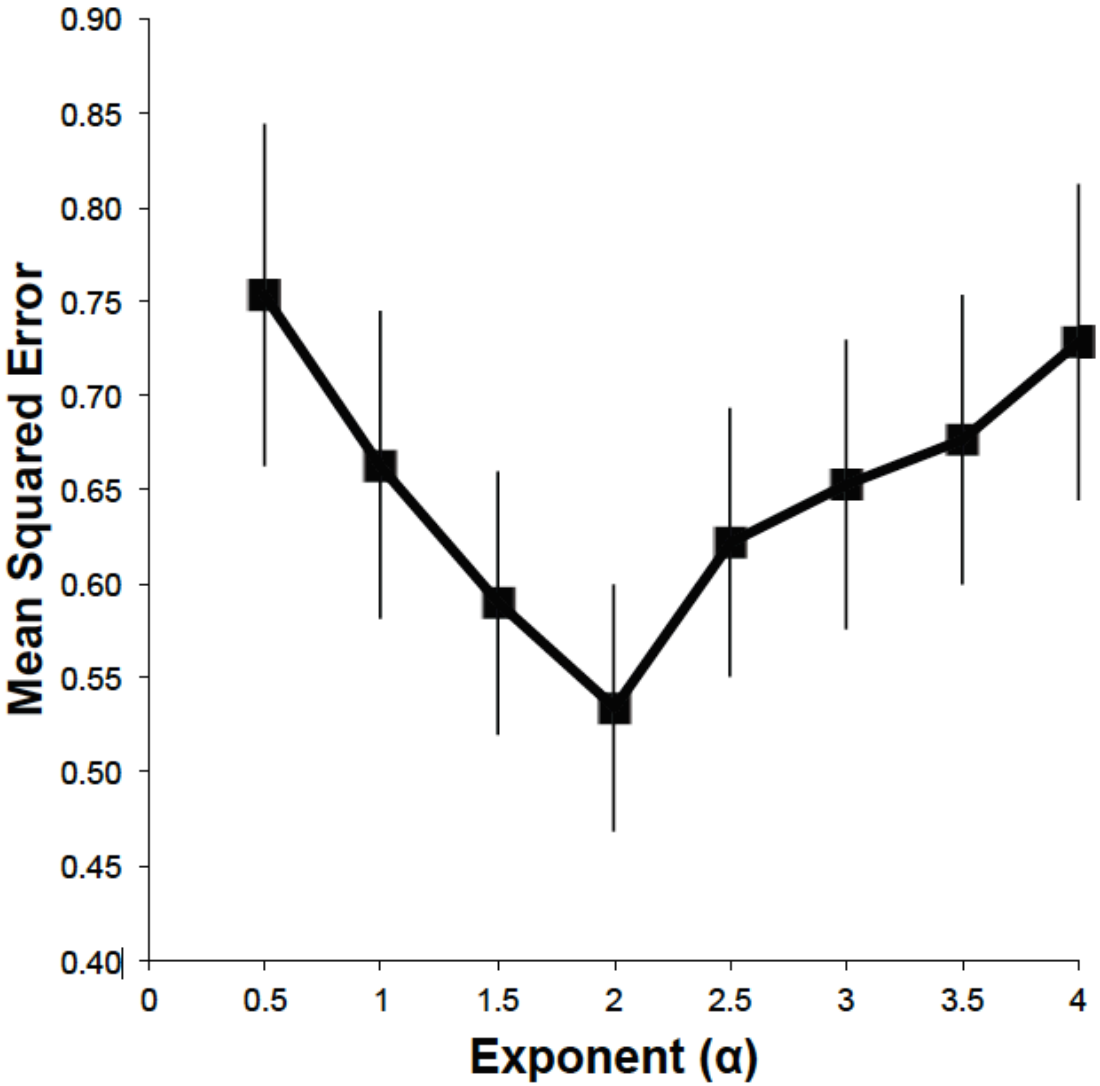


Figure 3 Leave-one-out cross-validation. Value of the mean-squared-errors as a function of α, obtained for a leave-one-out cross-validation on the HB-1119 dataset. The minimum error is found for an exponent of 2.0, although values of α between 1 and 3 do not produce significantly worse errors.

C (Distance Scaling Factor) Calibration

Without physical measurement of the distance between pairs of points along the sequence, it is difficult to accurately estimate the value of C . However, based on the average IF value of pairs of fragments located less than 5kb apart along the sequence and following Bystricky et al . [51] that packed chromatin has a physical length of 1 nm for every 110-150bp, C was estimated as approximately 50 nm.

Experiment

- Figure 4 shows that mixing is achieved after approximately 350 * 10^5 iterations, which requires less than 250 seconds of running time. Passed this point, structures sampled every 10^6 steps from the two parallel runs are undistinguishable from each other and sample structures from the same distribution.
- 250 structures were sampled after burn-in from each of the two runs. The two ensembles of structures were then combined and the 500 structures were clustered based on their structural similarity
- Analysis of the two THP-1 5C datasets produced similar results, and runs of a larger number of parallel MCMC chains confirm that they all sample similar structures.

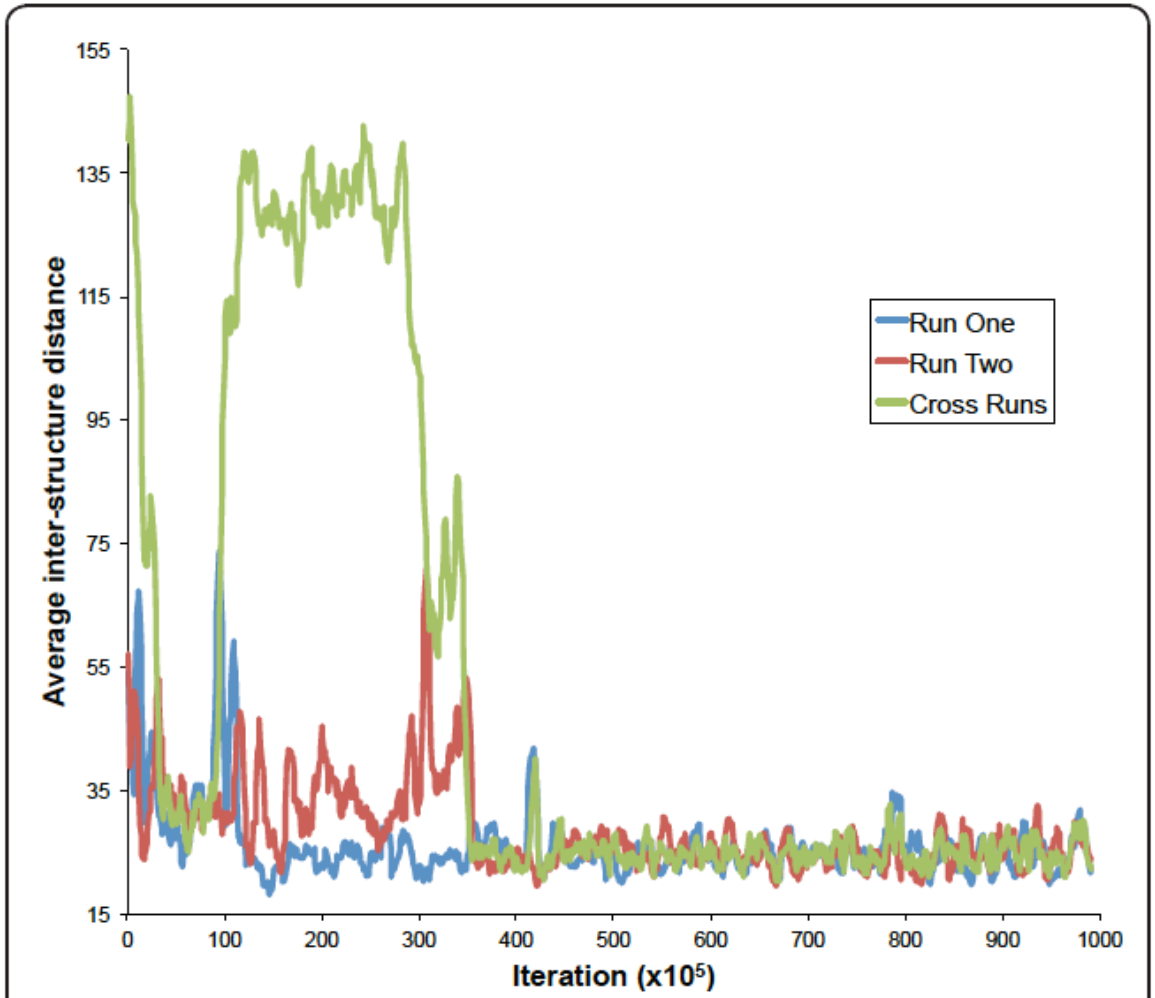


Figure 4 Mixing of parallel *MCMC5C* runs (HB-1119 dataset). Distance between consecutive structures (sampled every 10^6 iterations) from within one of two parallel *MCMC5C* runs (blue and red curves) or across the two runs (green curve), on the HB-1119 5C dataset. The runs converge to the same distribution very rapidly (in less than 250 seconds) and the cross-run distance (green) drops to within the same range as the within-run distances (blue and red curves) after 350×10^5 iterations.

Simulation Verification

Gold structure: a computationally constructed 3D structure used to generated IF data.

Simulated structure: models constructed from the IF data of the gold structure.

Verification by Sampling from Simulated True Structures

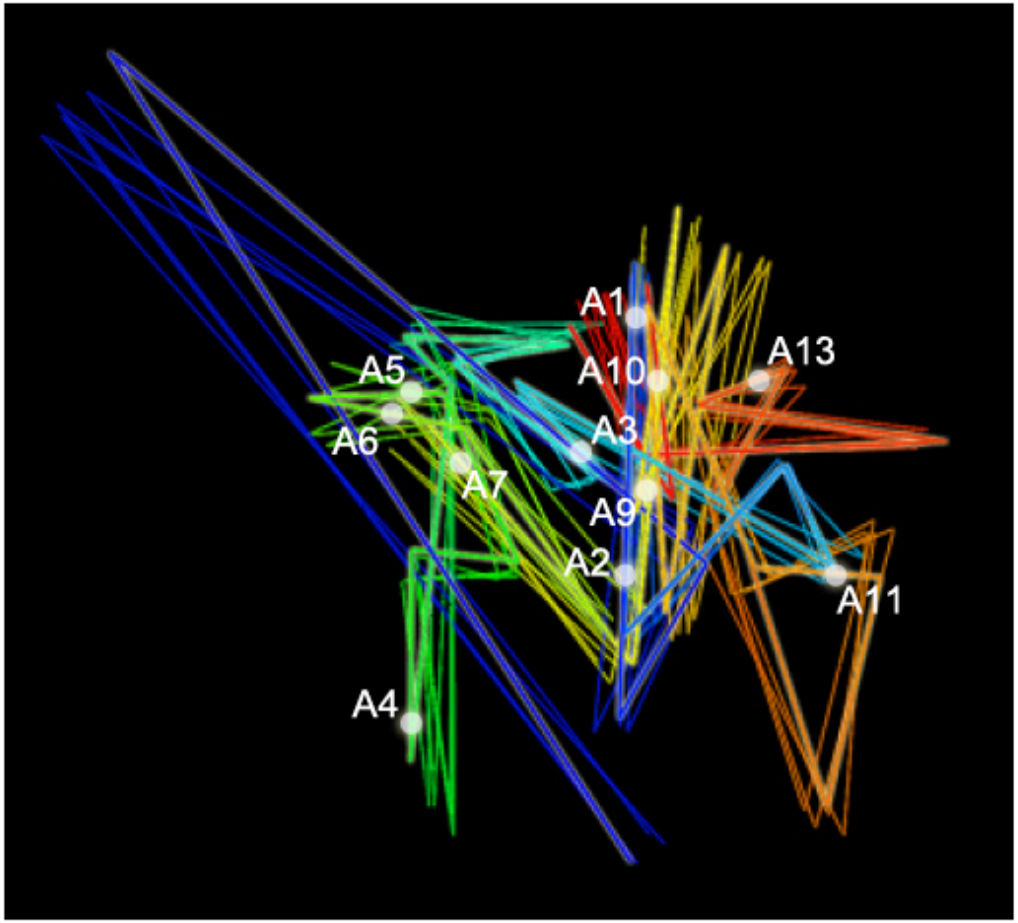


Figure 6 HB-1119 Structures from simulated data aligned to gold standard structure. The "gold standard" structure is used as a reference structure to which structures from four different parallel *MCMC5C* runs on simulated data generated from the gold standard structure are aligned. The gold standard structure is shown highlighted with a white glow and the transcription start sites for the HoxA genes are annotated. The structures found from the simulated data are shown in superimposition to the gold standard structure and show a high degree of alignment.

Structure Clustering and Sub-Structure Families.

Sub-structure families may correspond to chromatin structures of cells in different stages

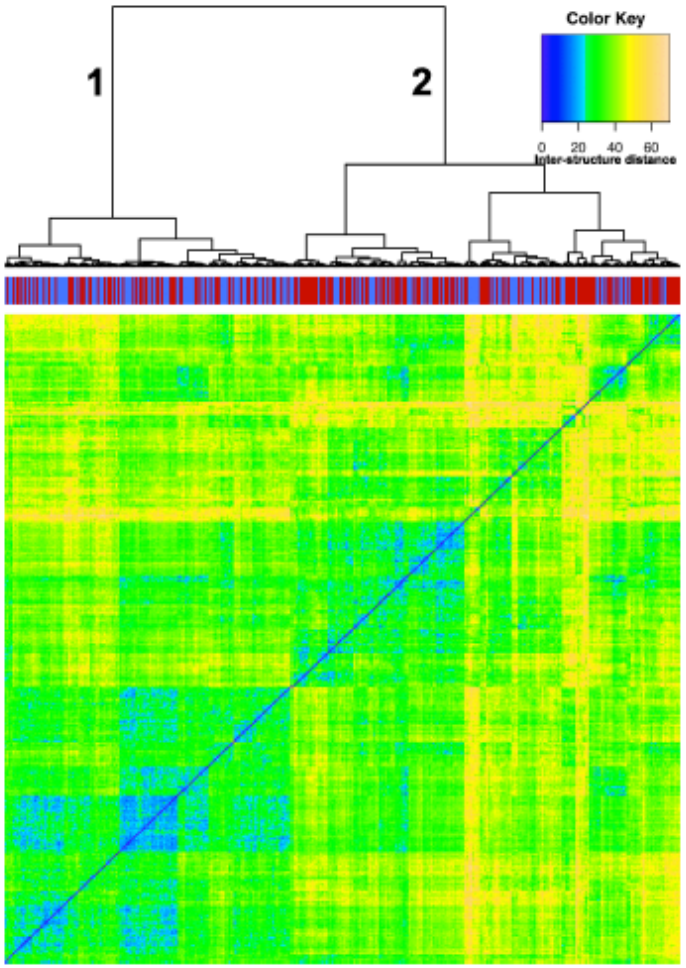


Figure 5 Mixing and subclustering of HB-1119 structures. Mixing and hierarchical clustering (Ward's method) of structure similarity. The five-hundred structures come from two parallel *MCMC5C* runs on the HB-1119 dataset (pools of 250 structures from each run were used). The colors along the top indicate which run each structure originated from (run one = blue, run two = red) and demonstrates that the sampling process has successfully mixed. The blocks in the heatmap and the dendrogram indicate the presence of sub-clusters of structures (numbered in the dendrogram). The two clusters (numbered 1 and 2) both contain structures from the two parallel runs (blue and red vertical bars), indicating that the structures are conserved across runs and are not an artifact of the burn-in process.

Conformations of HoxA in Undifferentiated and Differentiated Conditions

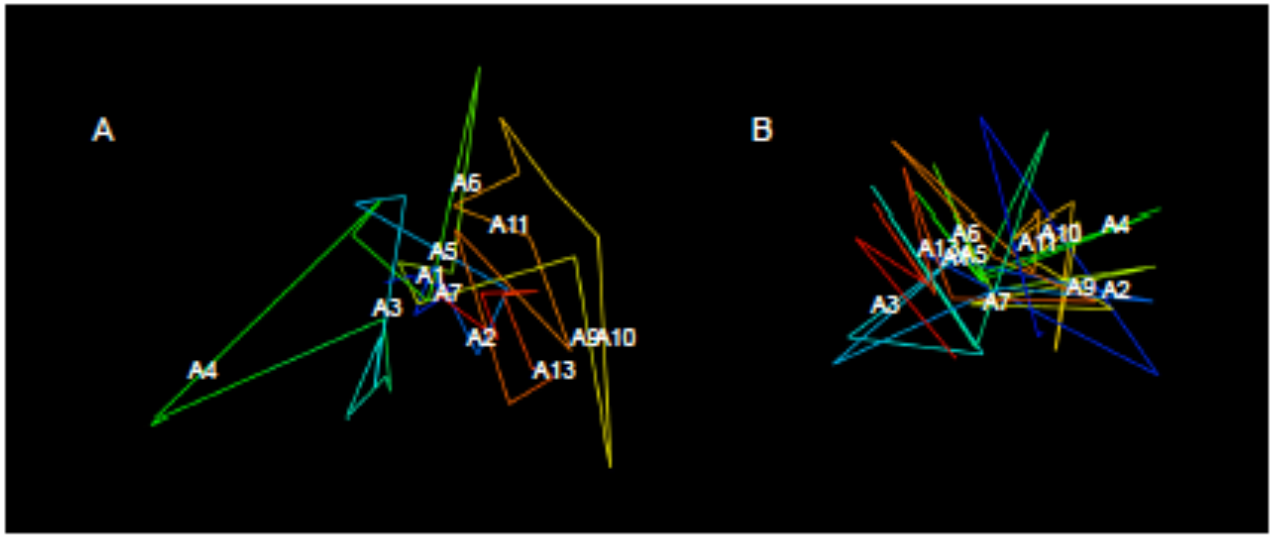


Figure 7 Models of HoxA cluster before and after differentiation. Maximum likelihood structures found by *MCMC5C* from the undifferentiated and differentiated THP-1 datasets (A and B, respectively). The HoxA gene transcription start sites are annotated on each of the structures.

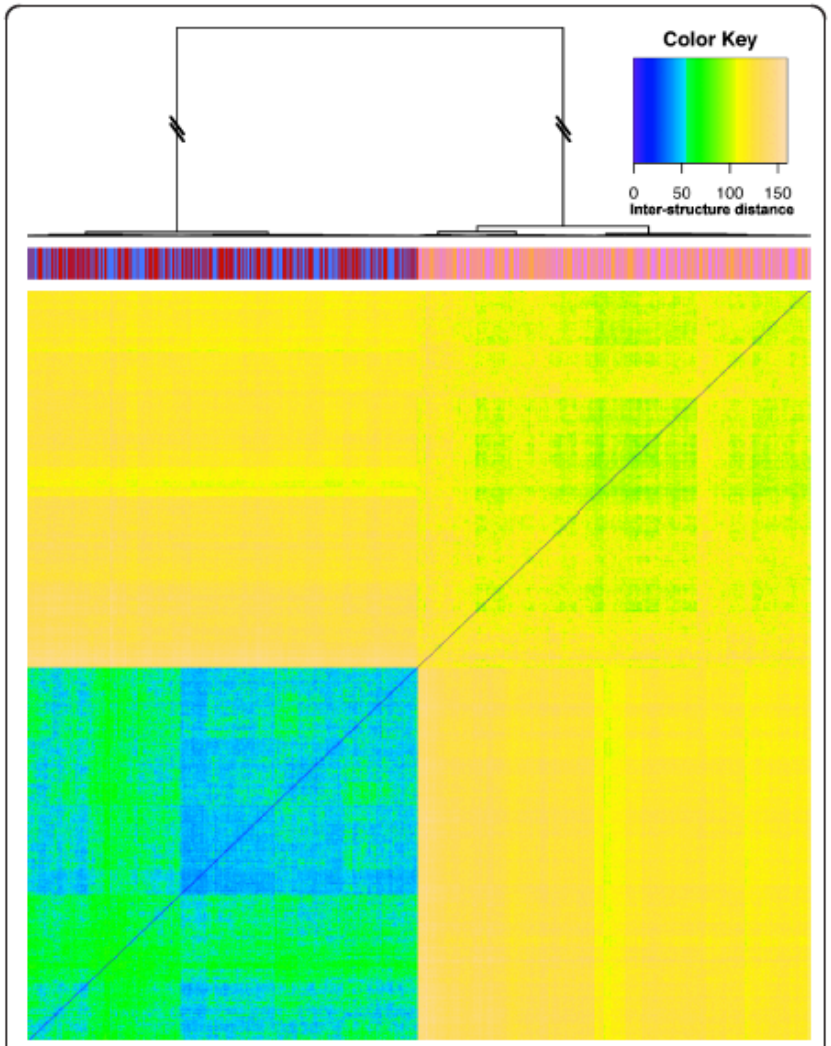


Figure 8 THP-1 clustering of undifferentiated and differentiated structures. Hierarchical clustering (Ward's method) of one-thousand structures from four parallel *MCMC5C* runs, two on the undifferentiated THP-1 dataset and two on the differentiated THP-1 dataset (250 structures each). The colors along the top indicate which state each structure originated from (undifferentiated run one = blue, run two = red; differentiated run one = pink, run two = orange) and demonstrate a clear distinction between the two states, indicating that the undifferentiated and differentiated cell states specify different structure signatures.

Structural Variation and Conservation

The subset of fragments that are the most conserved across the ensemble of structures are found to lie within the central core region of the structures. These fragments are spatially close to each other and may be involved in looping contacts that are important for the maintenance of the chromatin structure and are therefore highly conserved.

Analysis of Structural Properties

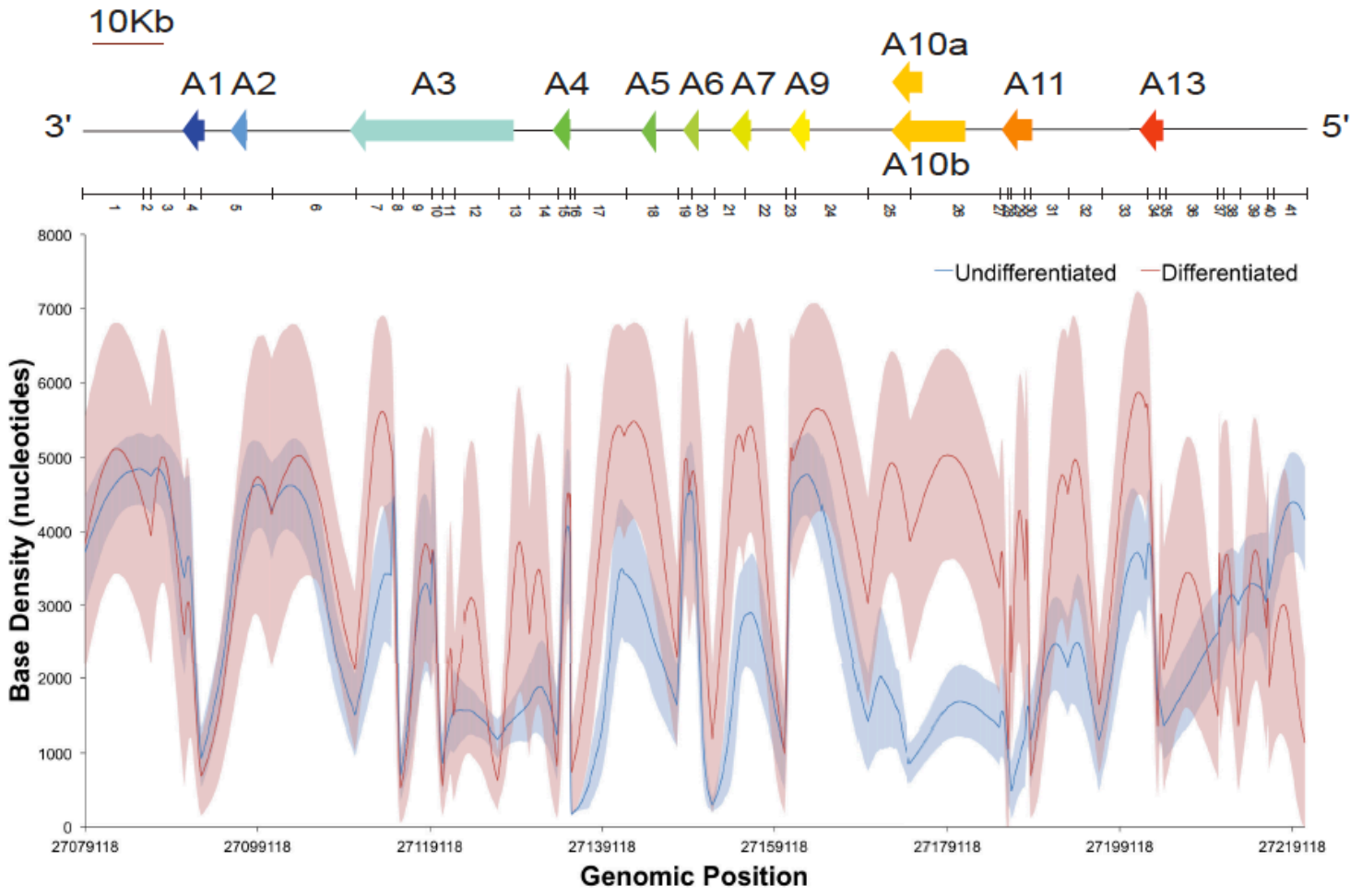


Figure 9 Base Density analysis of undifferentiated and differentiated THP-1 cells. Analysis of base density comparing the undifferentiated (red curve) and differentiated (blue curve) cell states. The genes in the HoxA locus are shown aligned above the plot. A pool of one-hundred structures generated by *MCMC5C* were used for each state. The base density measure was calculated with a sphere of radius one (1.0) every tenth base. The error bars report the standard deviation.

On Hi-C Data (Data Set II)

Model the long arm of human chromosome 14 (88.4 Mb region) from Hi-C data published by Lieberman-Aiden et al . [18] at a 1Mb resolution (89 fragments in total). Lieberman- Aiden et al . [18] proposed the existence of two physically disjoint compartments, whereby compartment A was found to correlate with open and actively transcribed chromatin, while compartment B was found to be more densely packed and repressed.

On Hi-C Data (Data Set II)

The authors designed four 3D-FISH probes (termed L1, L2, L3, and L4) that lie consecutively along chromosome 14 but alternate between compartments (A: L1 and L3; B: L2 and L4) and showed that the non-consecutive regions of the chromosome that belong to the same compartment appear to be physically closer than those that do not [18].

On Hi-C Data (Data Set II)

Our results using MCMC5C weakly supports this hypothesis, with the 3D-FISH probes L2 and L4 indeed being in close proximity. Importantly, we used an ensemble of 250 structures to estimate the distribution of predicted Euclidean distances between each pair of probes and found an excellent linear correlation with the physical distances measured by Lieberman-Aiden et al .

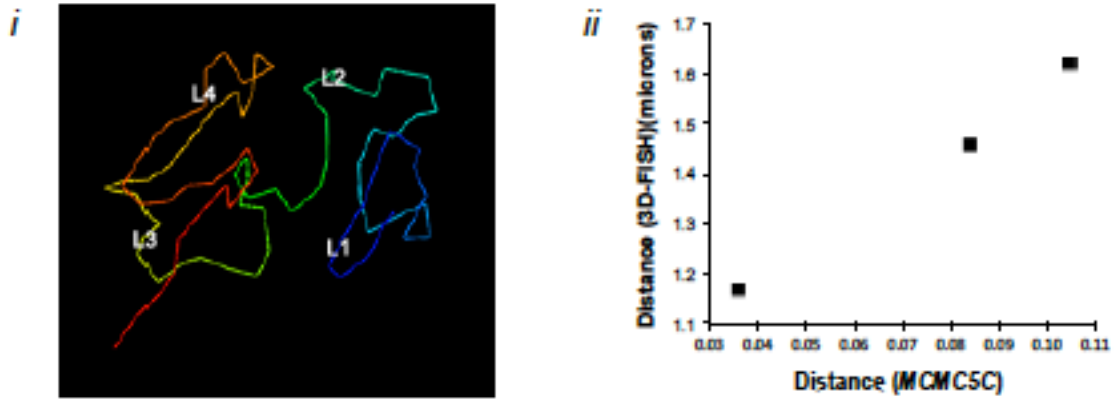


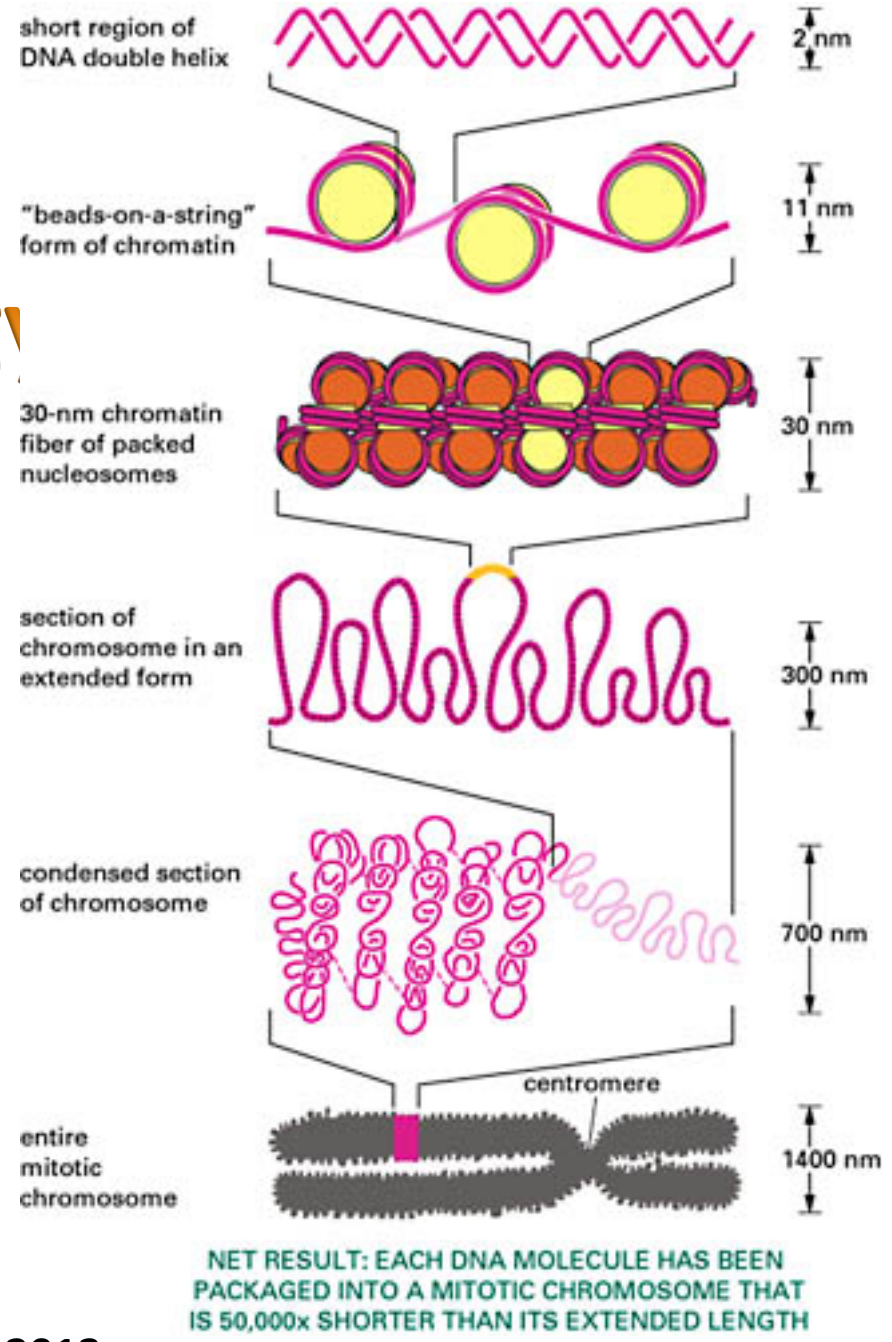
Figure 10 Modeling of human chromosome 14. (Left) *MCMC5C* model of human chromosome 14 from the Hi-C dataset. The midpoints of the 3D-FISH probes used by Lieberman-Aiden *et al.* [18] are annotated as L1, L2, L3, L4 and were designed such that in consecutive order the probes alternate between compartments. The structure adopts a loosely defined spiral form which brings the probes from within either compartment (A: L1 and L3, B: L2 and L4) in closer physical proximity than between pairs of probes across compartments. (Right) Distances inferred by *MCMC5C* correspond to physically-measured distances. X-axis: average Euclidean distance in the ensemble of 250 structures sampled by *MCMC5C*. Y-axis: median 3D-FISH physical distance measured by Lieberman-Aiden *et al.* [18]. Even though probe L3 is located between probes L2 and L4 in the linear sequence, probes L2-L4 are closer together in the model than L3-L2, indicating preferential organization of probes belonging to the same compartment (B) than across compartments (A-B) as initially reported in Lieberman-Aiden *et al.* [18].

MCMC5C Availability

- MCMC5C is available at http://Dostielab.biochem.mcgill.ca.

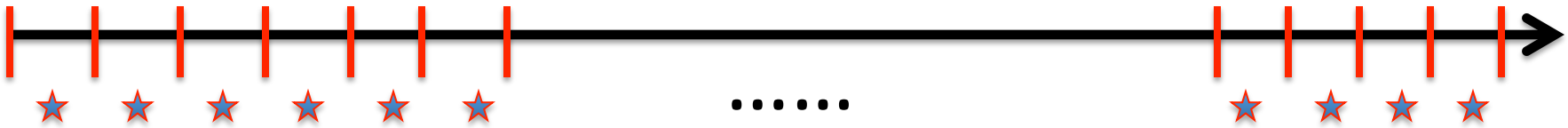
A Multi-Scale Modeling and Visualization Strategy

- Nucleosome (100b)
- Chromatin Fiber (Kb)
- Gene Loci (Mb)
- Chromosome (100Mb)
- Genome (Gb)



J. Cheng, NSF CAREER Project Plan, 2011, 2012, 2013

Spatial Representation of A Genome Region at a Scale



- A genome region (e.g. a chromosome) is divided into N equal / variable size units sequentially
- The spatial position of the center of a unit is denoted by a point and its coordinate (x, y, z) and constraints on the size of unit (radius of sphere)
- The consecutive points are joined into fragments forming the folding trace of the region.

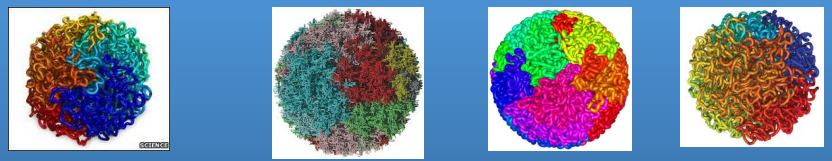
J. Cheng, NSF CAREER Project Plan, 2011, 2012, 2013

Structure Construction at One Scale

Initial Structure Representation of Units

Contact Map Between Units

Sampling & Optimization



J. Cheng, NSF CAREER Project Plan, 2011, 2012, 2013

Contact-Driven Modeling at Chromosome Scale

- **Input**: initial representation of chromosome, contact map, and physical distance restraints
- **Objective**: find 3D chromosome structures that satisfy the contact map and physical distance restraints as much as possible.
- **Scoring Function**
- **Optimization**
- **Output**: an ensemble of 3D shapes

J. Cheng, NSF CAREER Project Plan, 2011, 2012, 2013

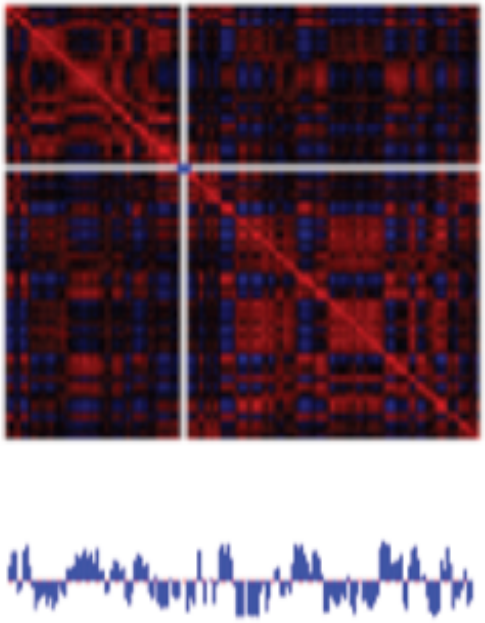
Structure Modeling Movie

At YouTube without music: http://www.youtube.com/watch?v=C03R7A9kYc8

At NSF CAREER project web site with music:
http://people.cs.missouri.edu/~chengji/genome_modeling_movie.mp4

J. Cheng, NSF CAREER Project Plan, 2011, 2012, 2013. T. Tuan made the movie.

Two Compartment Validation on Normal B-Cell

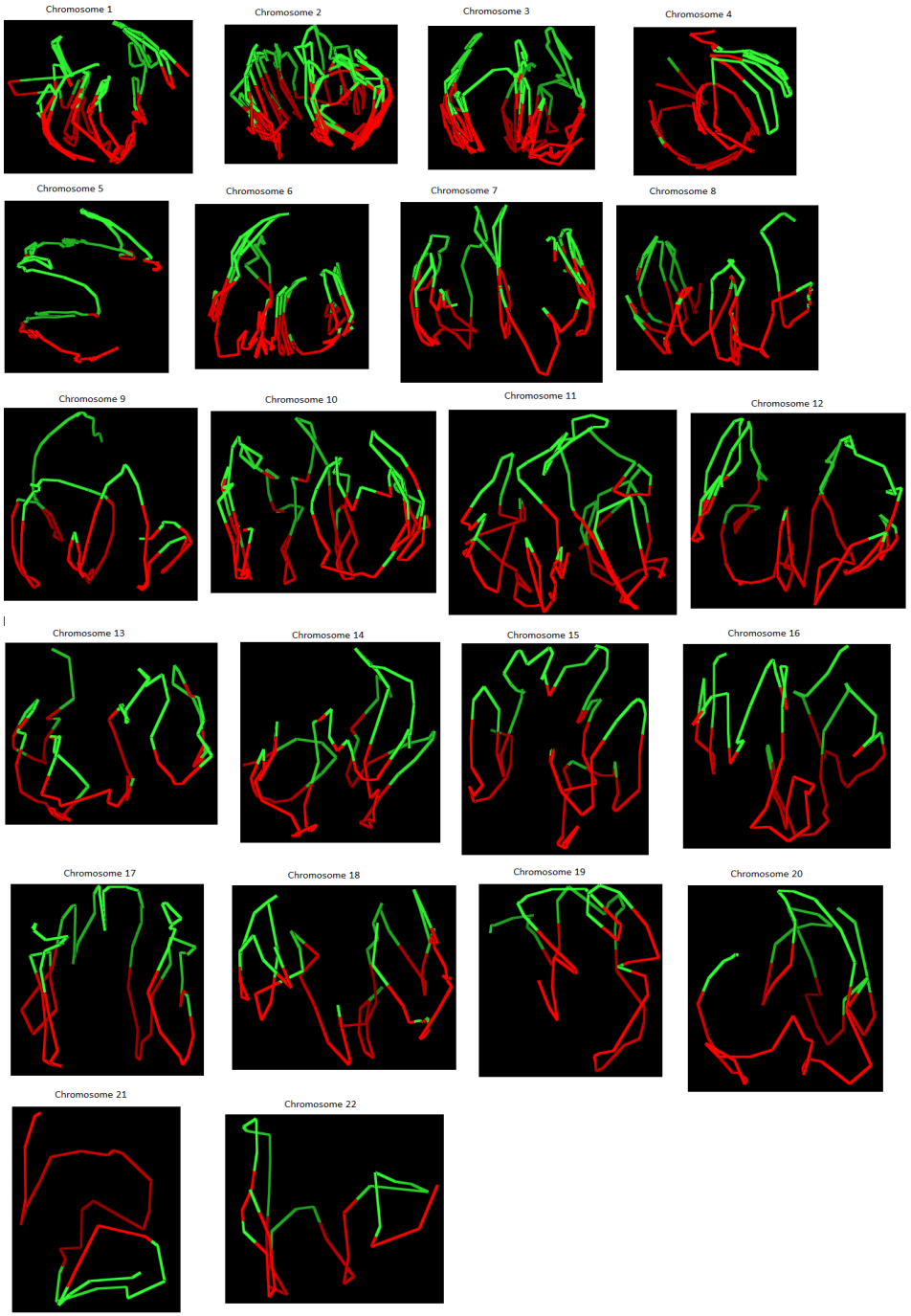


Chromosome 7

Purple and green denote regions in two different components using principle component analysis on contact correlation map

J. Cheng, NSF CAREER Project Plan, 2011, 2012, 2013

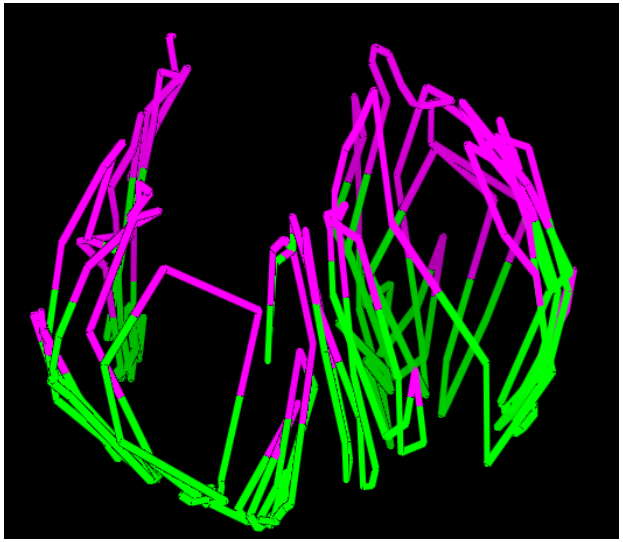
3D Models for 22 Chromosomes of Normal B-Cell



J. Cheng, NSF CAREER Project Plan, 2011, 2012, 2013

Models of Chr. 2 for Normal and Malignant Cells

Primary Leukemia B-Cell



Normal B-Cell

J. Cheng, NSF CAREER Project Plan, 2011, 2012, 2013

Demo: GMol - Multi-Scale Visualization of 3D Genome Structure

Acknowledgements

- Zheng Wang, Renzhi Cao
- Tuan Trieu, Chenfeng He, Sharif Ahmed
- Avery Wells
- Charles Caldwell, Kristen Taylor, Aaron Briley