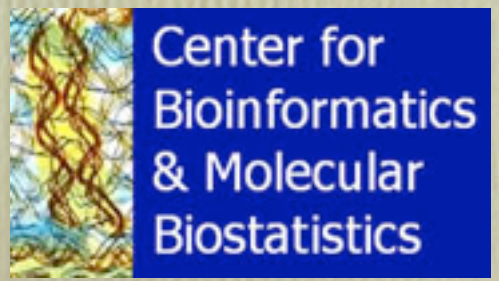
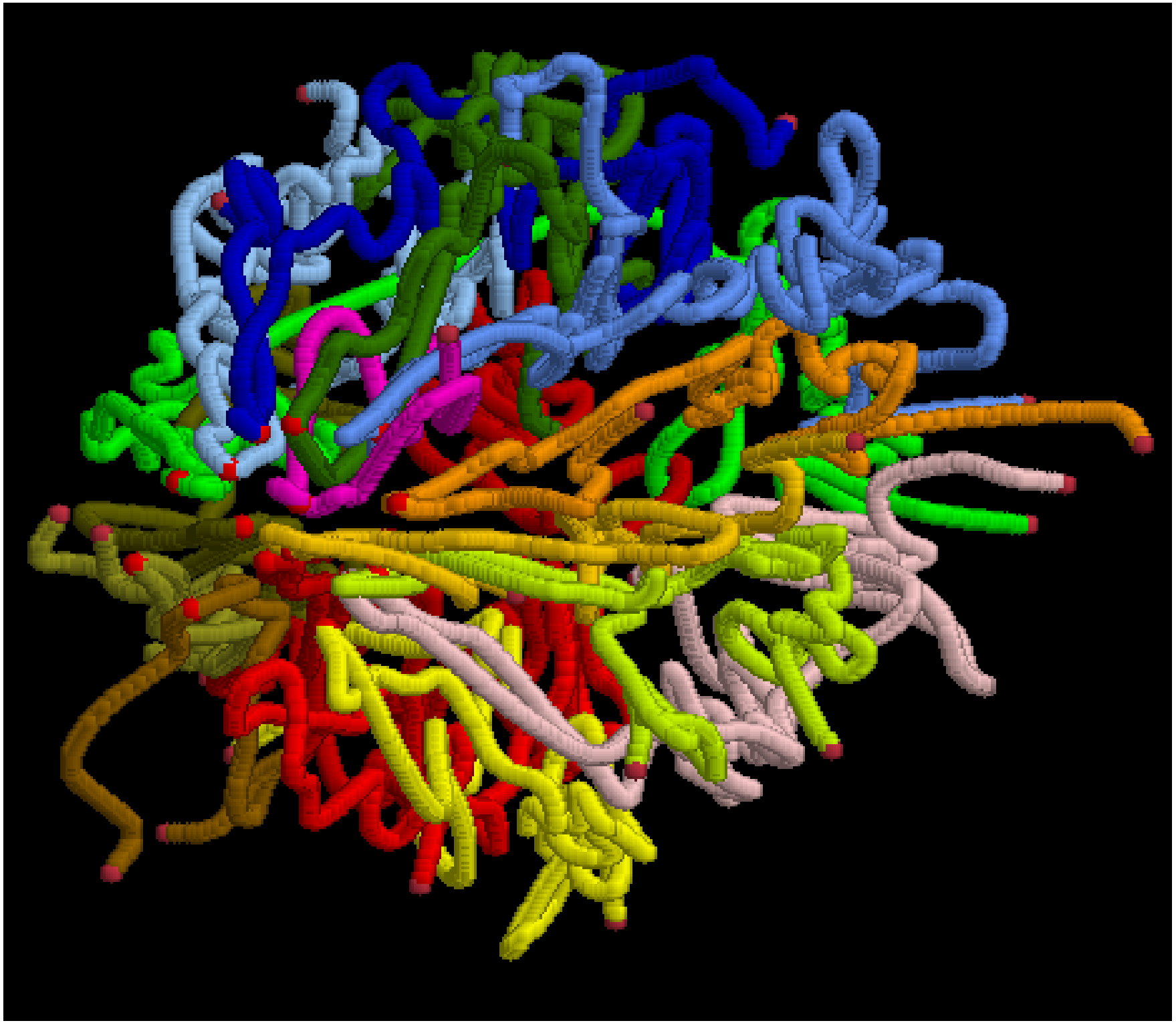
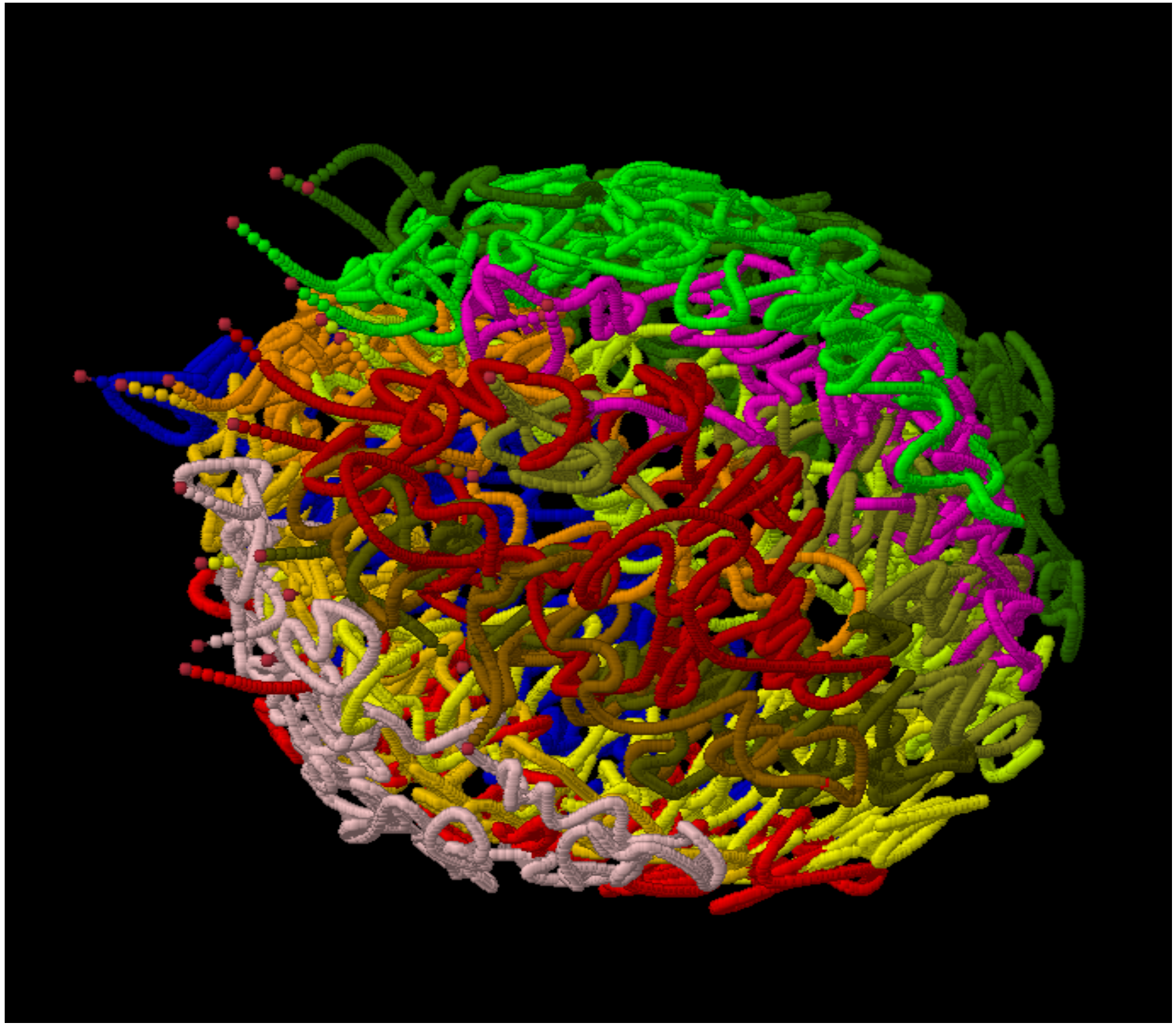
Reconstructing 3-D Genome Configurations: How and Why

Mark Segal
Center for Bioinformatics & Molecular Biostatistics
UCSF Divisions of Bioinformatics &Biostatistics

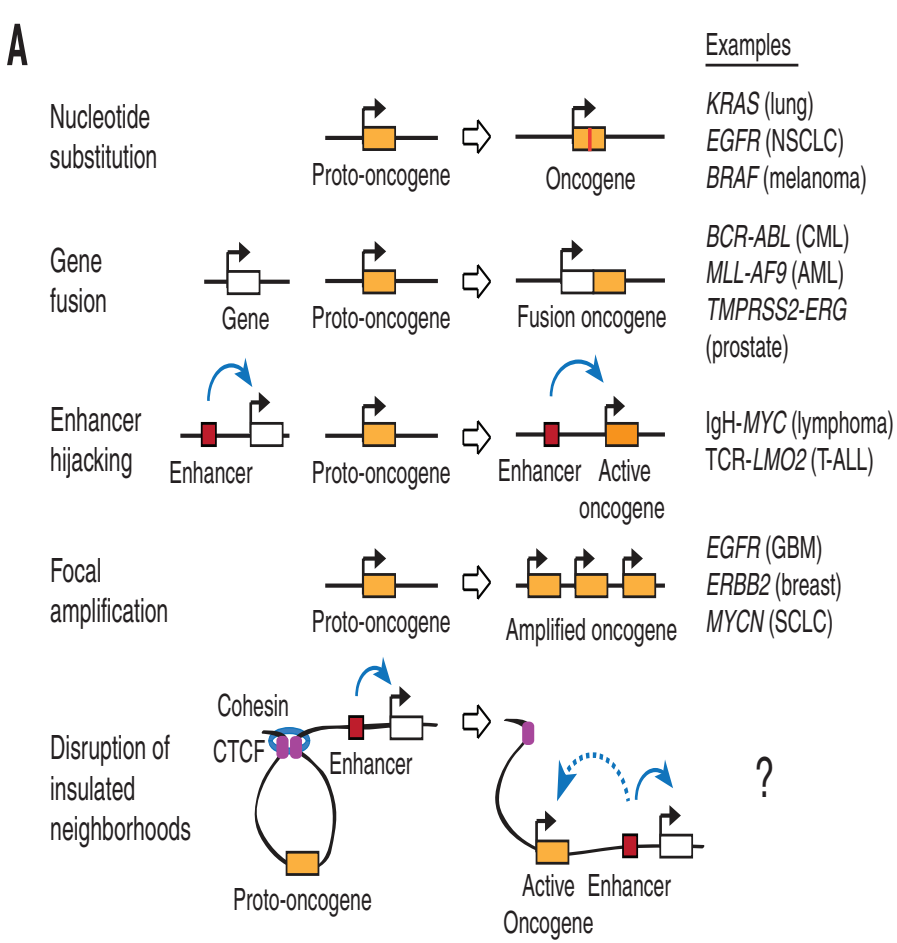


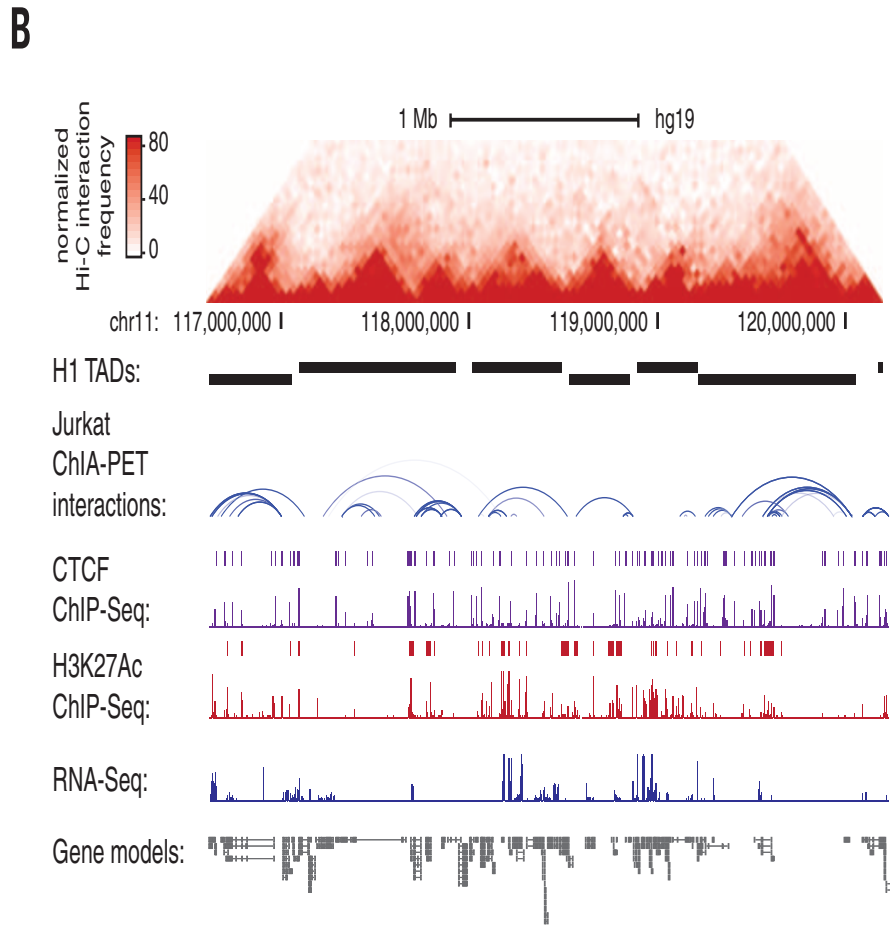
BMI206 - Statistical Methods for Bioinformatics
November 21, 2016





Oncogenesis via neighborhood disruption





Outline

- Why 3D genome architecture (theory)
- Chromatin conformation capture assays
- How 3D structure is inferred:
  - algorithm choices and issues
  - reproducibility / accuracy assessment
- Why 3D genome architecture (practice)
- Further possibilities

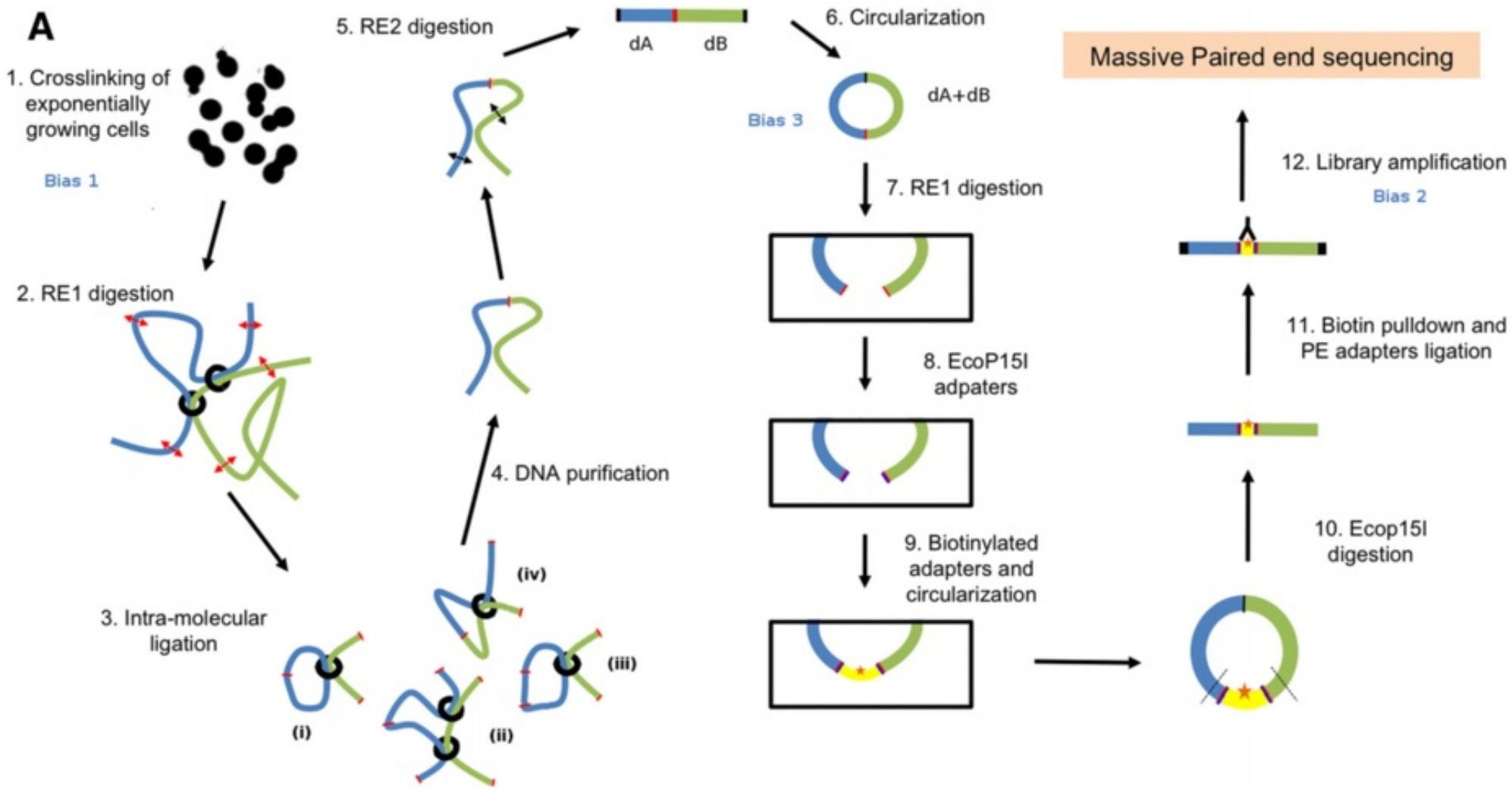
Importance of 3D Architecture

- Gene regulation:
  - co-localization of co-expressed genes into transcription factories
  - positioning of distal control elements
- Translocations / gene fusions:
  - 20% of human cancer morbidity
  - 3D structure “probably pivotal”

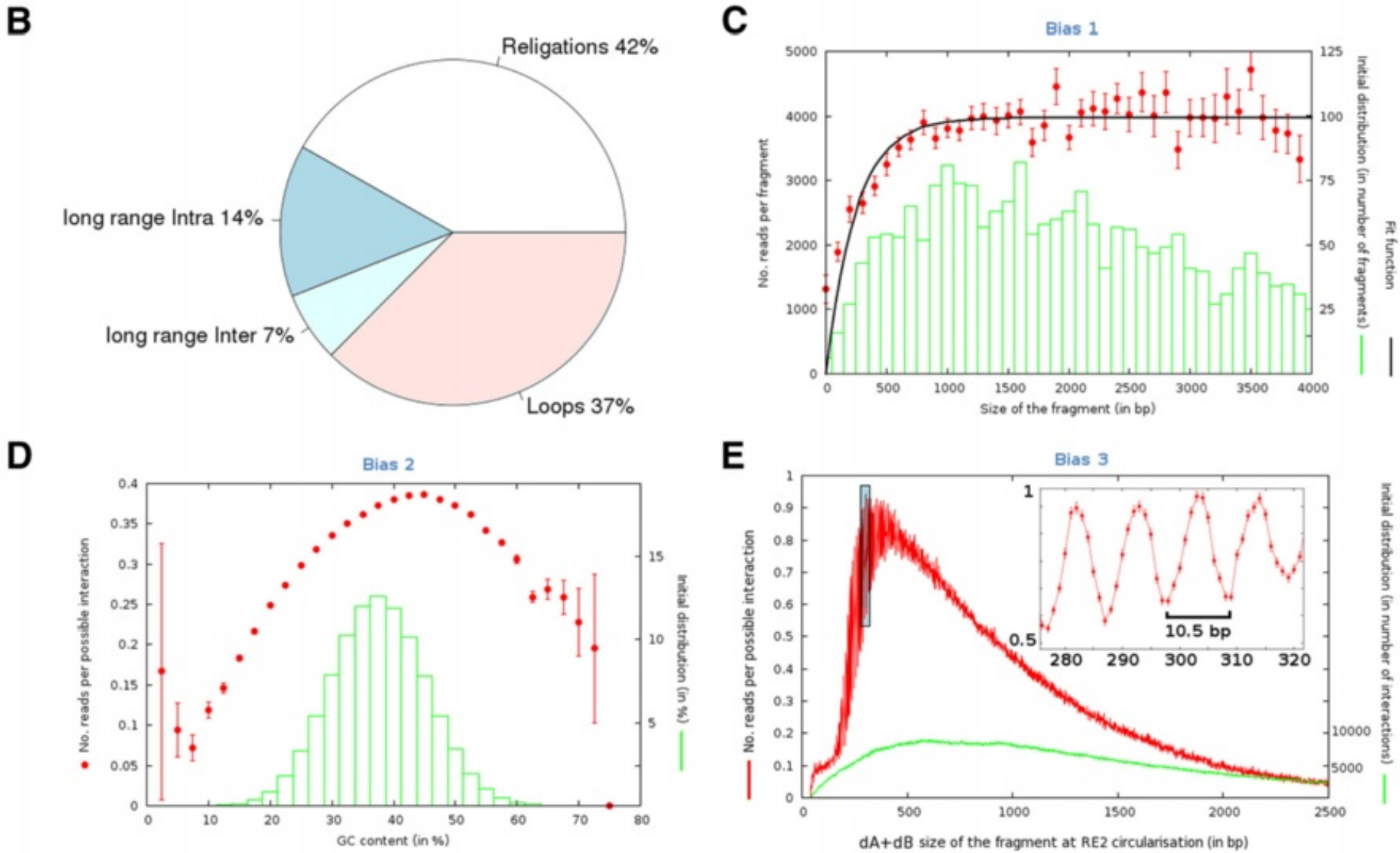
Observing / Inferring 3D Structure

- Challenging at even modest resolutions:
  - genomes are highly condensed
  - genomes are dynamic, variable
  - traditional assays are low throughput and low resolution (FISH coarse)
- Recently devised suite of Chromatin Conformation Capture techniques has revolutionized 3D structure elicitation

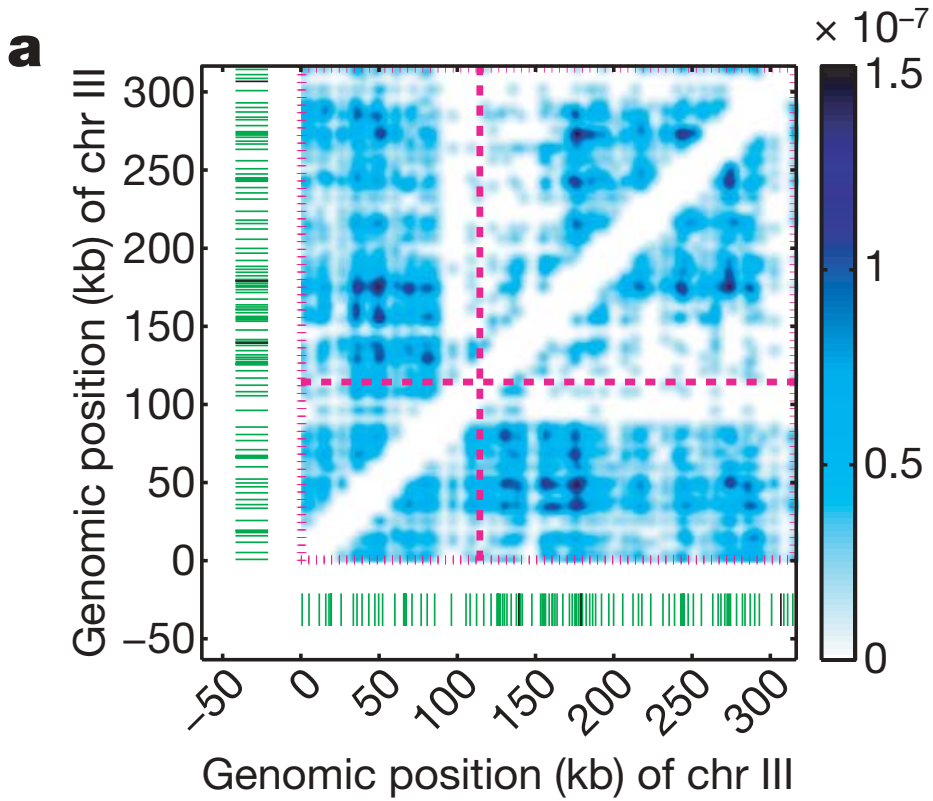
3C / 4C / 5C / Hi-C / TCC

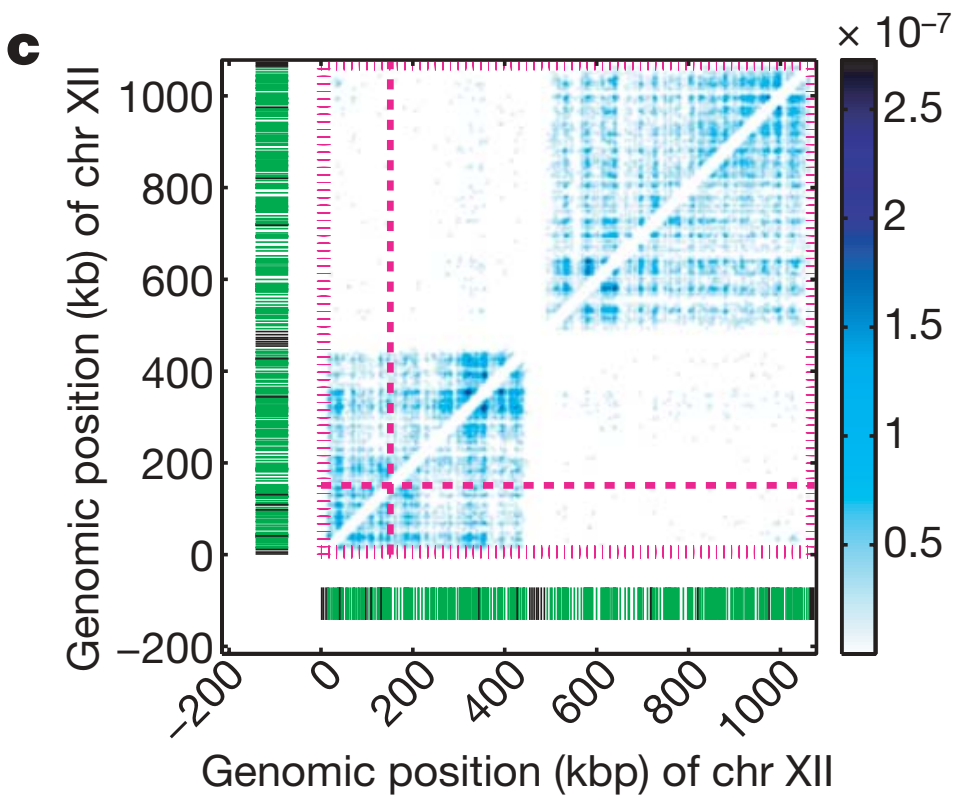


Bias Correction / Normalization



Output: Contact / Interaction Maps





Also *inter*-chromosomal maps.

Assume *m* total loci after possible binning.

From Interactions to 3D Structure

- Objective: given an interaction matrix *F*, obtain a 3D structure (or an ensemble thereof) the between-loci pairwise distances of which are highly correlated with the corresponding interaction frequencies in *F*.
- Two broad classes of approach:
  - Optimization / consensus procedures
  - Ensemble / probabilistic procedures

Ensemble / Probabilistic Methods

- Ensemble motivation: assay performed on hundreds of thousands of cells -- single structure summary is misleading; providing a collection of solutions displays genomic structural variation.
- This reasoning is entirely aspirational: there is no basis for equating displayed variation to biology -- could be purely algorithmic.
- Much downstream analysis will require a single structure -- back to consensus.

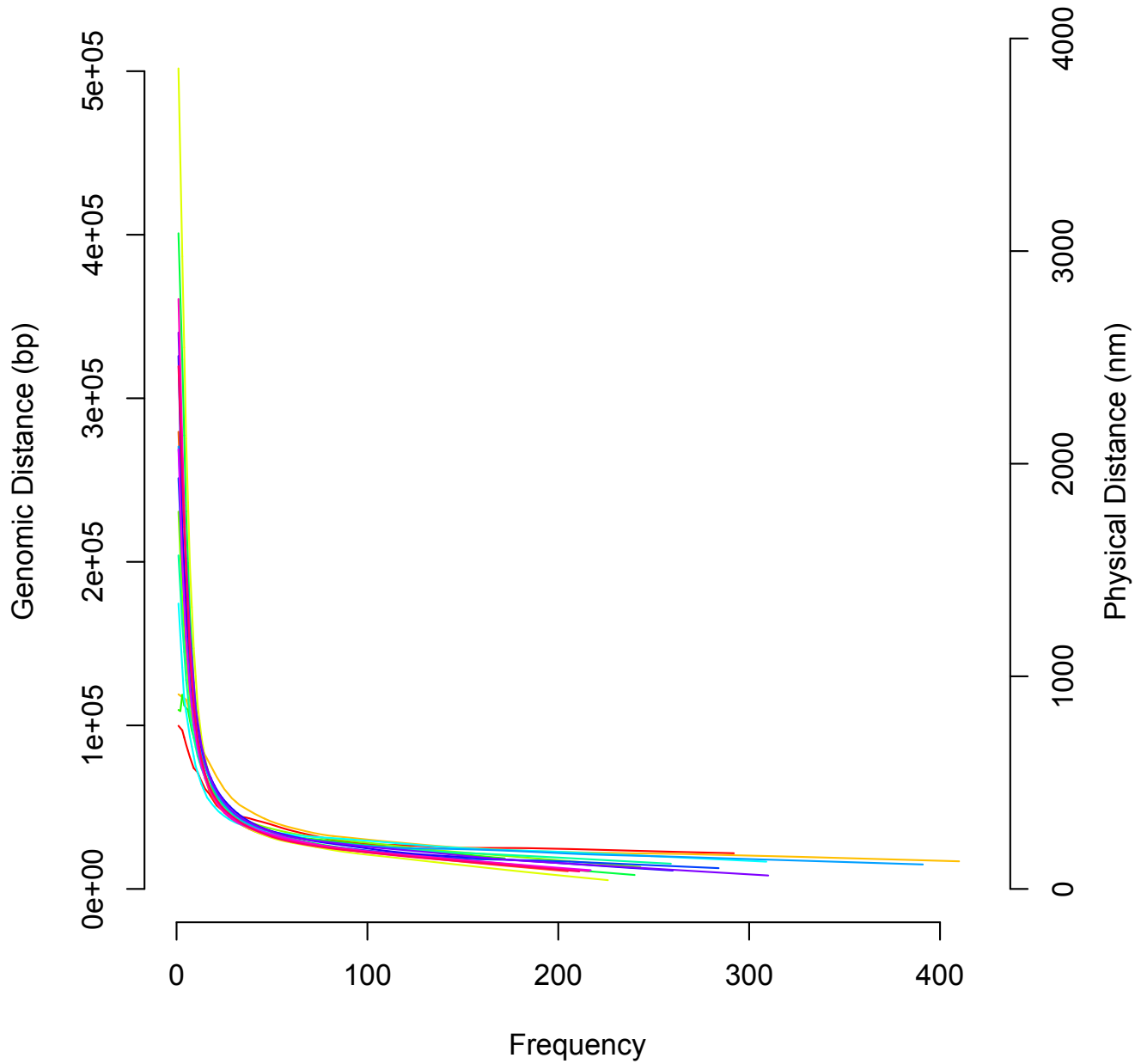
Optimization / Consensus Methods

- Generally utilize two steps:
  - convert *F* into a distance matrix *D* that captures expected pairwise distances
    - differing strategies / assumptions
    - sometimes interplay with second step
  - learn / estimate 3D structure from *D*
    - multi-dimensional scaling (MDS)
    - weights, non-metric variants

Interactions to Distances I

- Can empirically relate intrachromosomal interactions to genomic distances
- *Saccharomyces cerevisiae*: every ~130 bp of packed chromatin has length 1 nm
- Provides a simple ruler for conversion of frequencies to physical distances
- Obtaining physical distances enables incorporation of biology based constraints into the subsequent MDS optimization step
- Duan et al *Nature* (2010)

Conversion to Physical Distances



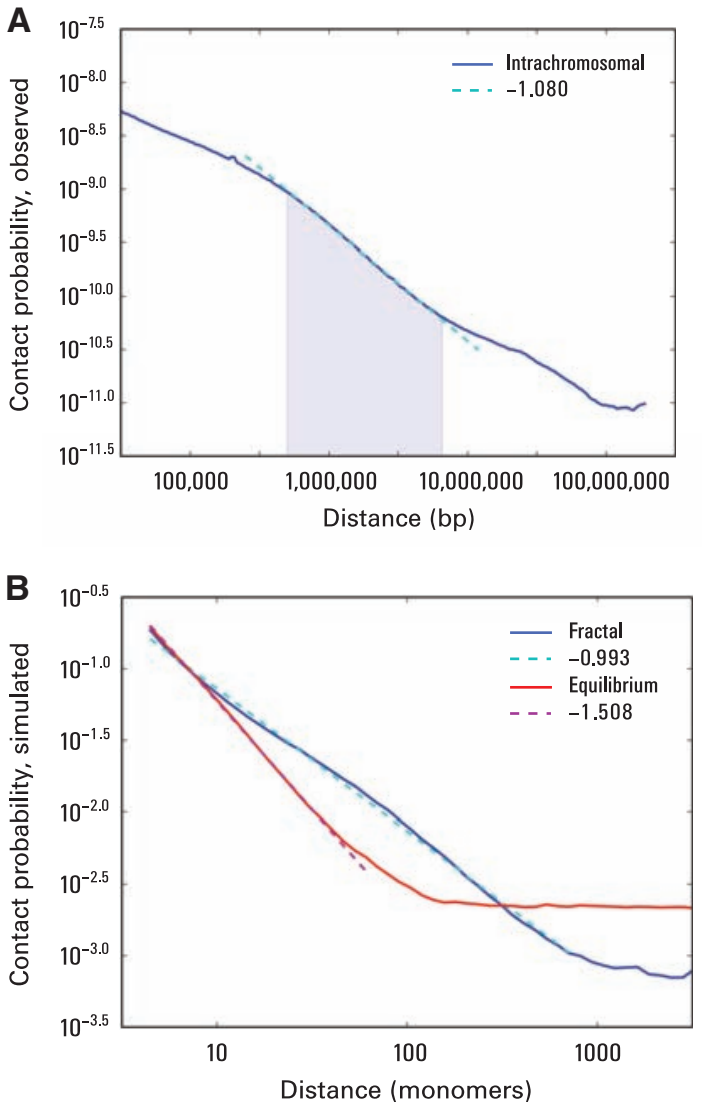
Interactions to Distances II

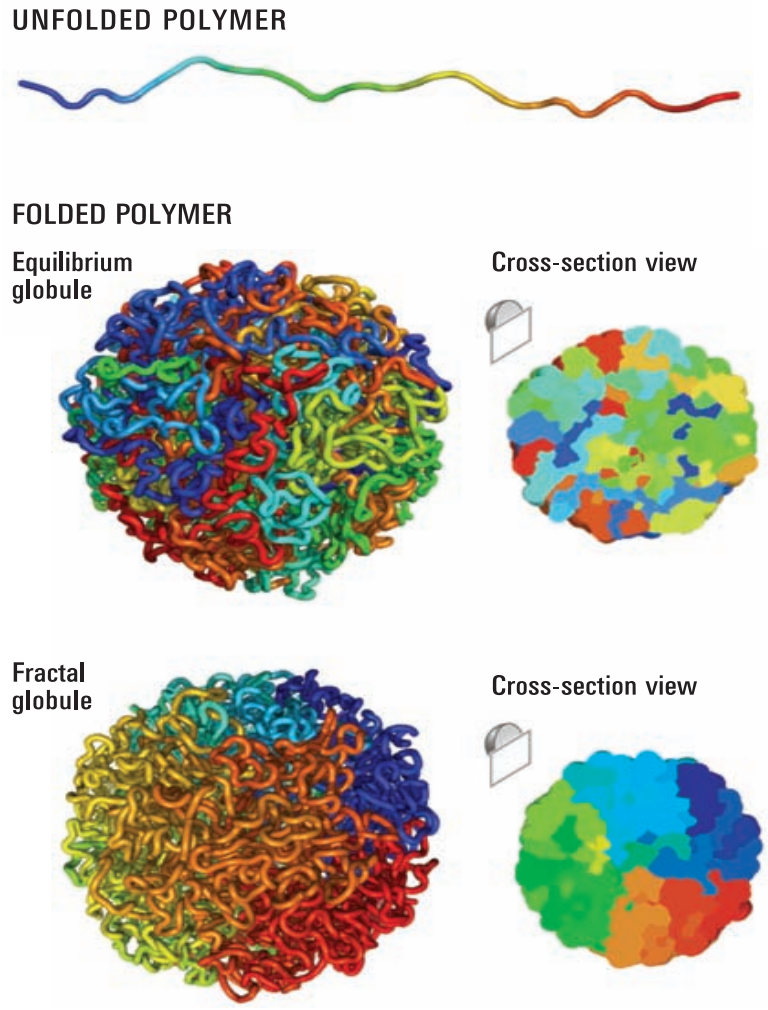
- Empiric & theoretic [polymer biophysics, fractal / equilibrium globules] results support power law relationship between *F* and *D*

$$D \propto F^{-\alpha};\ \alpha > 0$$

- But index thereof can vary according to organism, resolution, cell cycle phase...
- Estimate index from data [*cf* NMDS]
- Zhang et al *J Comp Bio* (2013) ChromSDE
  Zou et al *Genome Biology* (2016) HSA

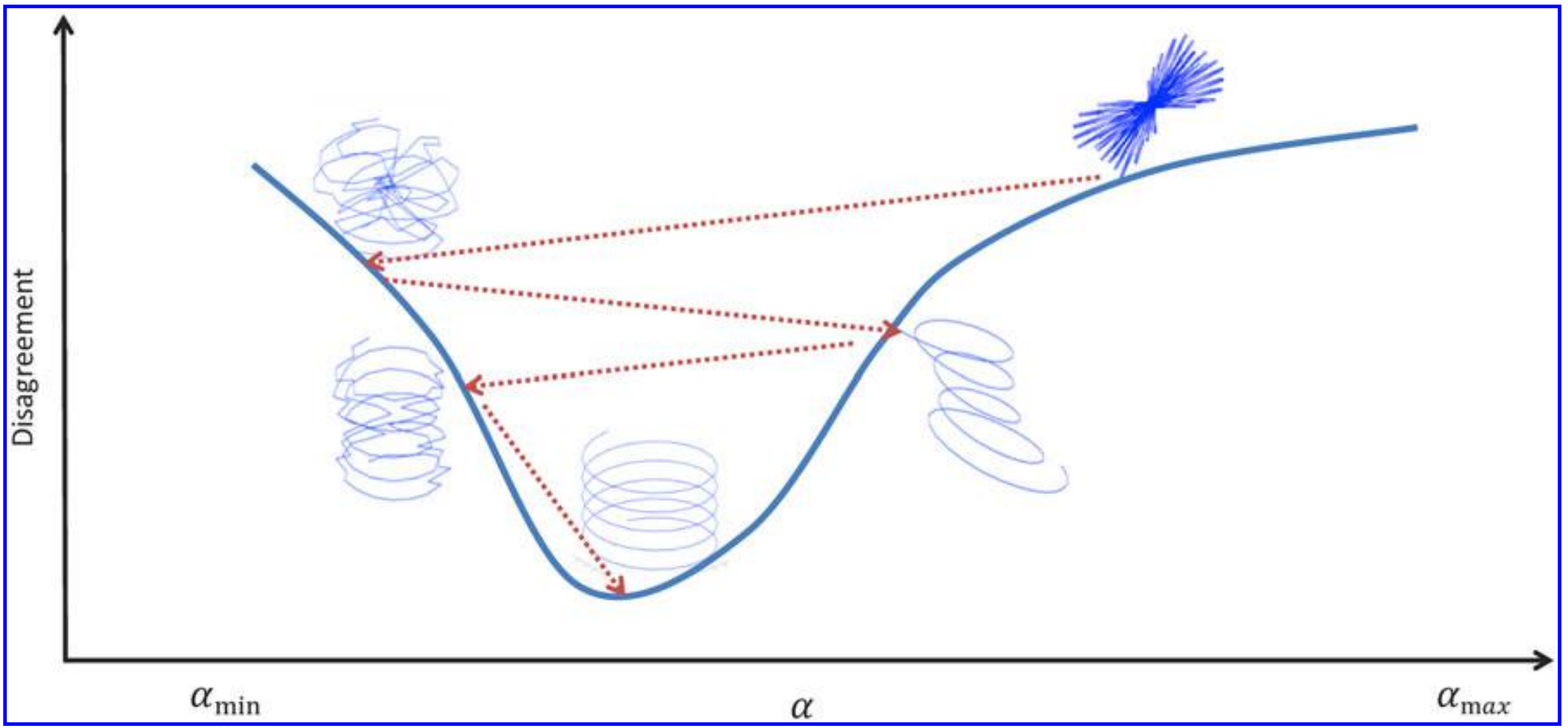
Basis for Power Law





- Lieberman-Aiden et al *Science* (2009)

ChromSDE Index Estimation



Interactions to Distances III

- Create weighted graph whose nodes are detected loci and length of link is inverse contact frequency
- Distance between loci is then length of the shortest path connecting them:
  - Computed using Floyd-Warshall algorithm
  - Can handle single cell Hi-C assays
  - Derived distance purportedly robust
- Lense et al *Nat Meth* (2014) ShRec3D

1D Distances to 3D Structure I

- Minimize objective function that places (as much as possible) interacting loci at their expected distance apart (MDS):

$$\min_{\{x_i, x_j \in R^3\}} \sum_{\{i,j|D_{ij}<\infty\}} \omega_{ij} \cdot (\|x_i - x_j\| - D_{ij})^2$$

- Benefit of obtaining physical distances *D* is provision for imposing biology based constraints.

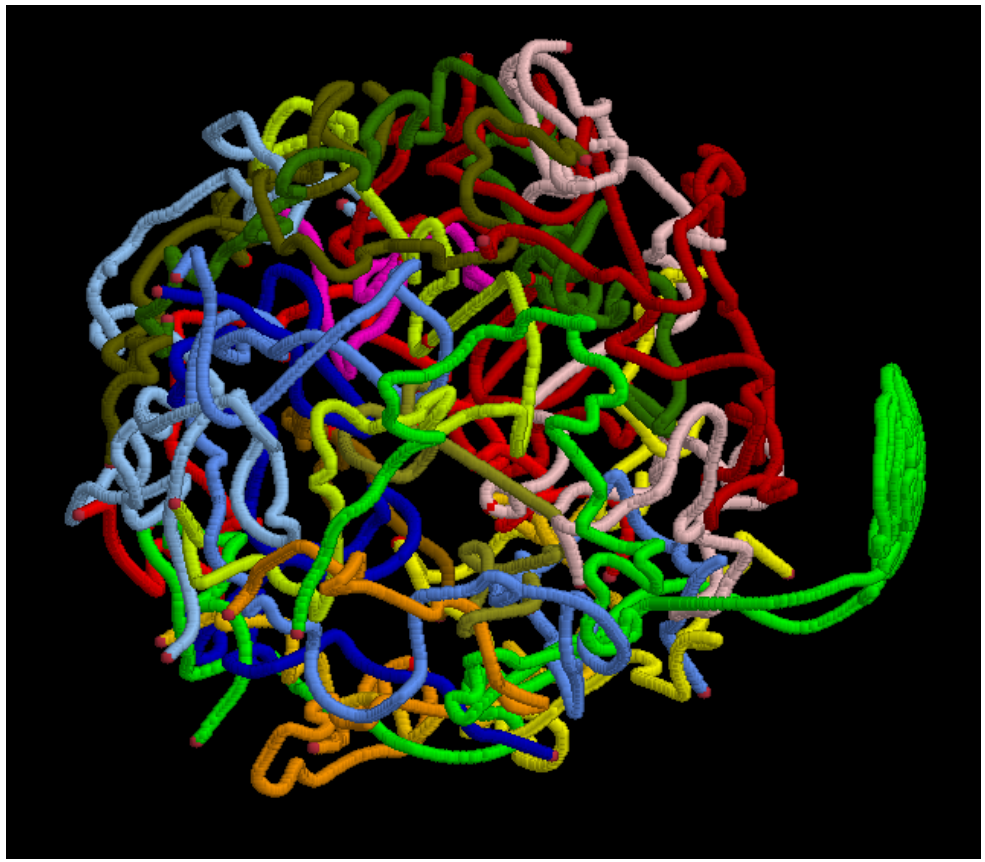
Physical Distance Constraints

- All points in 1μm sphere (yeast nucleus).
- Adjacent points within a given range.
- No two points on same chromosome can be closer than 30nm (chromatin fiber).
- Minimum distance between points on different chromosomes.
- rDNA repeats within the nucleolus.
- Centromeres cluster opposite nucleolus.

1D Distances to 3D Structure I

- Drawbacks to using physical distances:
  - Strong assumptions to obtain ruler
  - Organism specific formulations
  - Slow, delicate (interior point) optimization:
    - with yeast loci spaced at 10kb there are x 10^3 parameters, 10^6 constraints
    - ~ 2.5 days to solve; not parallelizable
    - Sensitivity to inputs, data: challenging

Structure Reproducibility



HindIII

EcoRI

Differing restriction libraries

Structure Reproducibility

$0.066^2 \leq d^2(p,q) \leq 0.091^2$

$0.066^2 \leq d^2(p,q) \leq 0.08^2$

Differing adjacency constraints

1D Distances to 3D Structure II

- Minimize objective function that places (as much as possible) interacting loci at their expected distance apart (MDS):

$$\min_{\{x_i, x_j \in R^3\}} \sum_{\{i,j | D_{ij} < \infty\}} \omega_{ij} \cdot (\|x_i - x_j\| - D_{ij})^2$$

- Penalty: $$\lambda \sum_{\{i,j | D_{ij} = \infty\}} \|x_i - x_j\|^2$$

  - Non-interacting loci cannot be too close

1D Distances to 3D Structure II

- Nonconvex, nonlinear optimization: NP hard
- Existing methods use heuristics to solve:
  - MCMC, Simulated annealing (IMP - Sali)
- By relaxing solution space from *R*^3 to *R*^*m* problem becomes convex semidefinite:
  - Global minimizer in polynomial time
  - Recover exact solution in noise-free setting
- Zhang et al *J Comp Bio* (2013) ChromSDE

1D Distances to 3D Structure II

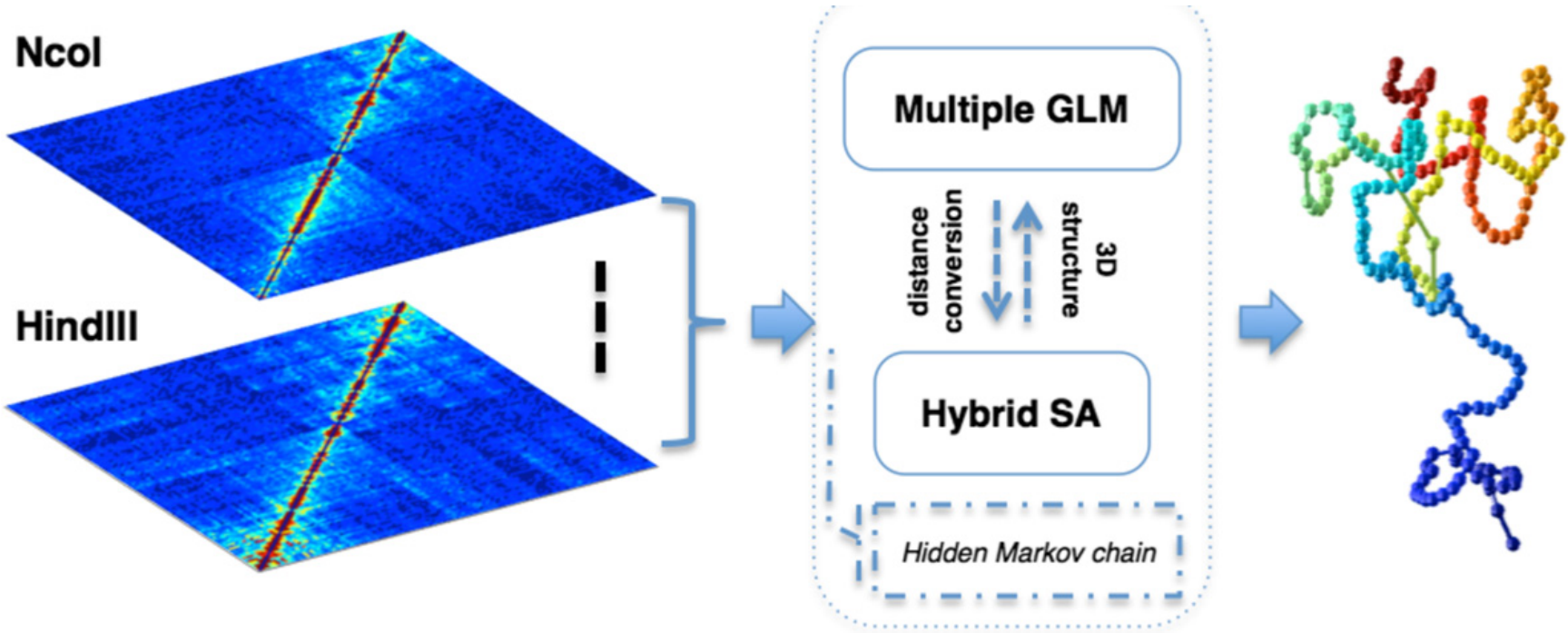
- But ... computational considerations limit problem size: number of loci / resolution
  - ChromSDE uses a sophisticated quadratic SDP solver that handles much larger problems ($m \sim 3000$) than general SDP solvers ($m \sim 200$)
  - Corresponds to 1 Mb resolution for human
  - Need 100 kb resolution to capture *topological domains*: highly self-interacting regions Dixon et al *Nature* (2012)

- This has resulted in
  - single chromosome solutions: no whole genome insights Varoquaux et al *Bioinformatics* (2014)
  - downsampling and/or simple organisms
- No 3D genomes for mouse, human
  - ShRec3D exception ...
- Note: New in situ (intact nuclei) assay has improved contact map resolution to ~ 1kb revealing new domains Rao et al *Cell* (2014)

1D Distances to 3D Structure III

- ShRec3D advantages include speed, problem size capacity
  - But no index estimation: ChromSDE, HSA
  - Distances ascribed to zero, small frequencies ostensibly filtered but criteria for such filtering unclear
    - Potentially big component of speed
- Purported insensitivity to index prescription

1D Distances to 3D Structure IV



- HSA advantages include handling multi-track data, use of starting configurations, built-in normalization

Two-Stage Hybrid Proposal

- Expand scope of existing methods to provide higher resolution, whole genome reconstructions
- Use ChromSDE or HSA per chromosome:
  - 3D coordinates using *intra*-chromosomal counts [bulk of counts ~ 15 - 20 fold]
  - Power law index
- Stitch together using *inter*-chromosomal counts [bulk of interacting pairs ~ 10 fold]

Sample (select?) n_k points from ChromSDE solution for chromosome k. Let $n = \sum n_k$

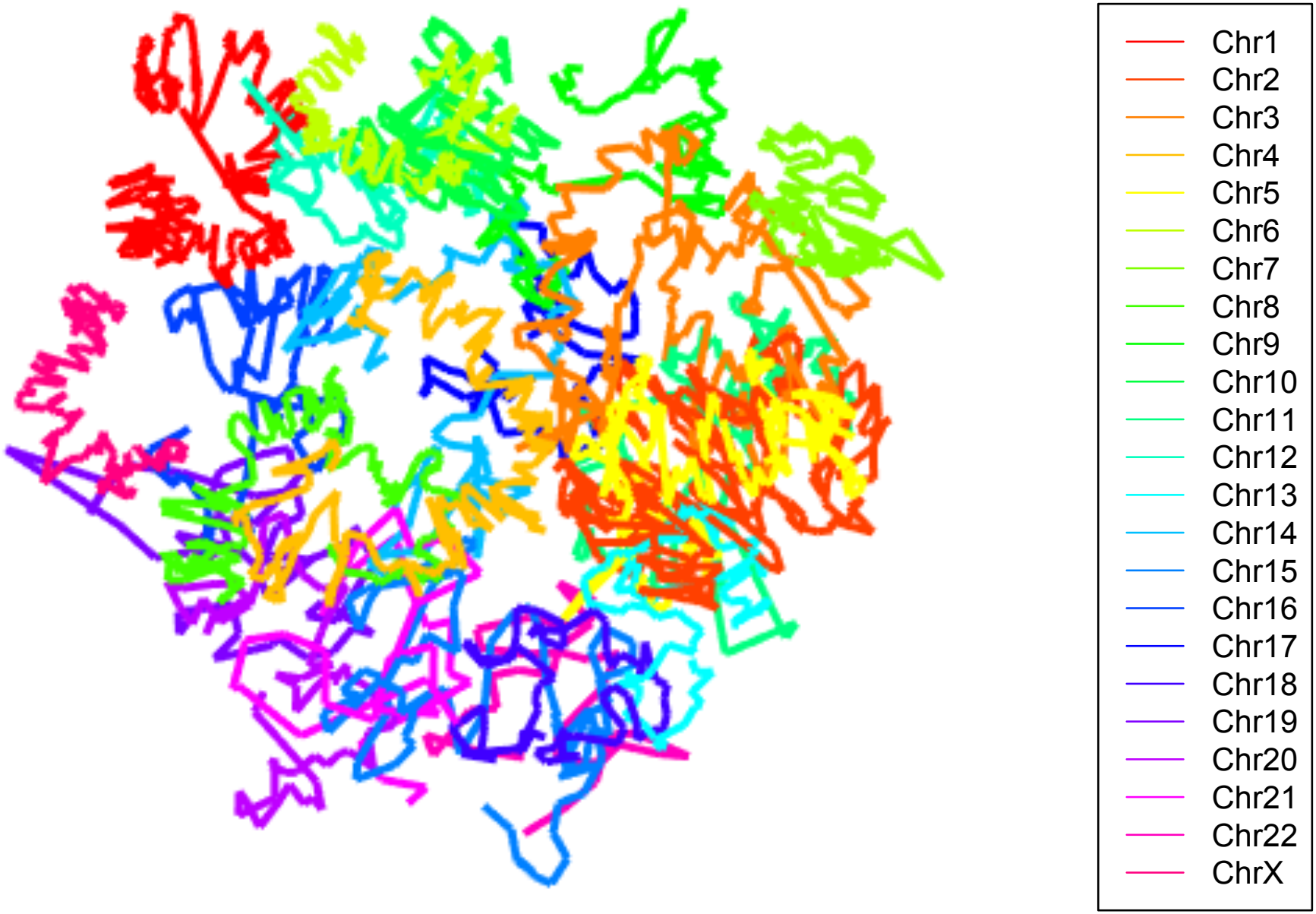
Intra-chromosomal inter-point distances obtained from 3D coordinates hybrid algorithm

Inter-chromosomal inter-point distances:
$D_{ij} = (F_{ij})^{-\sqrt{\alpha_k \cdot \alpha_{k'}}} \quad i \in \mathrm{Chr}_k, j \in \mathrm{Chr}_{k'}$

Configuration based on $D_{n \times n}$ via (N)MDS

Map original ChromSDE solutions to the configuration via Procrustes transformation

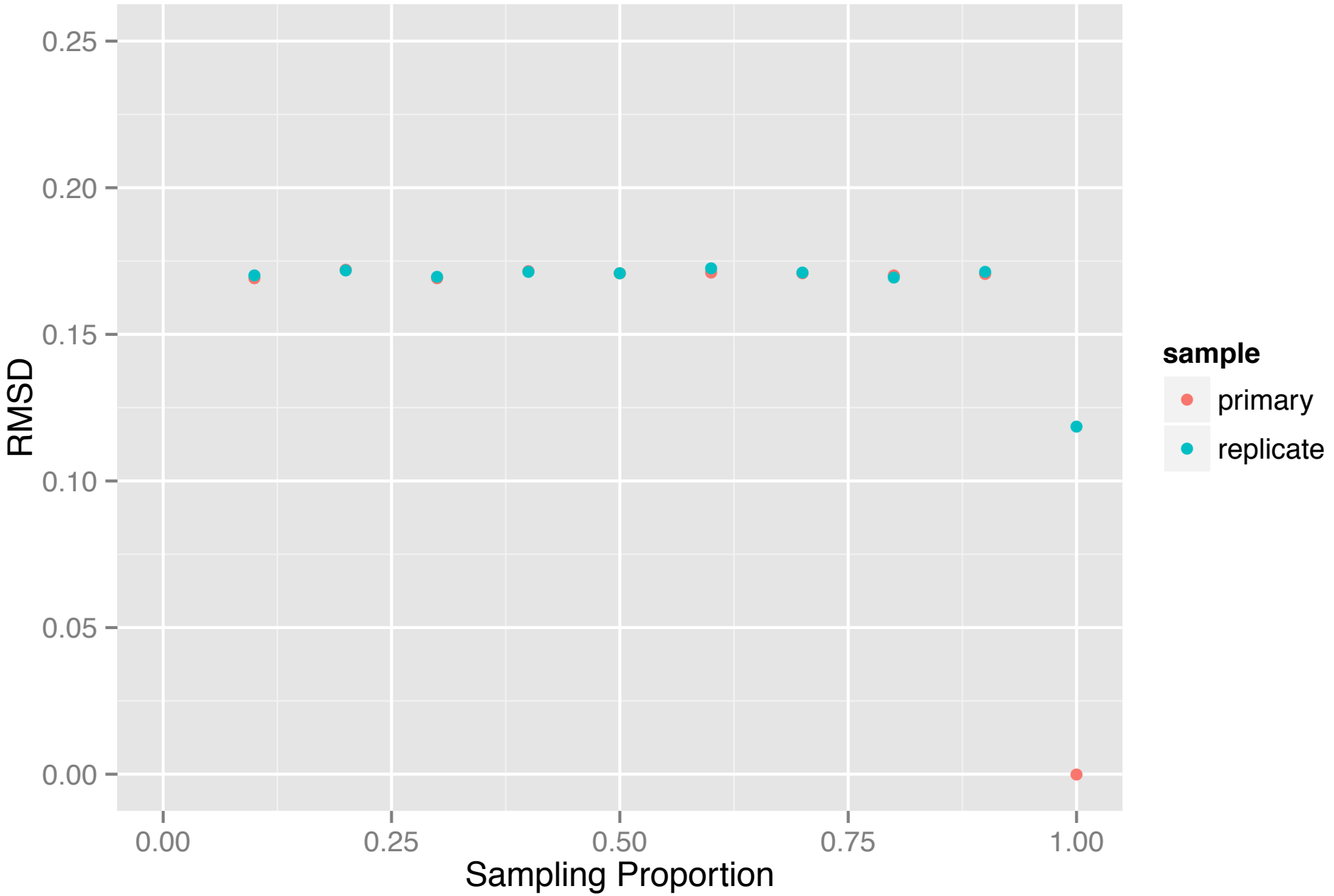
Lymphoblast Reconstruction



Evaluating Sampling Impact

- No real handle on accuracy:
  - known properties, FISH landmarks
    - crude**; only individual chromosomes
- Assessing reproducibility also difficult:
  - Differing REs construed as replicates
  - Inference: permutation, null-referent dns Segal et al *Biostatistics* (2014)
  - In situ Hi-C provides genuine replicates for a range of cell-lines and resolutions

Reproducibility: In situ 1Mb

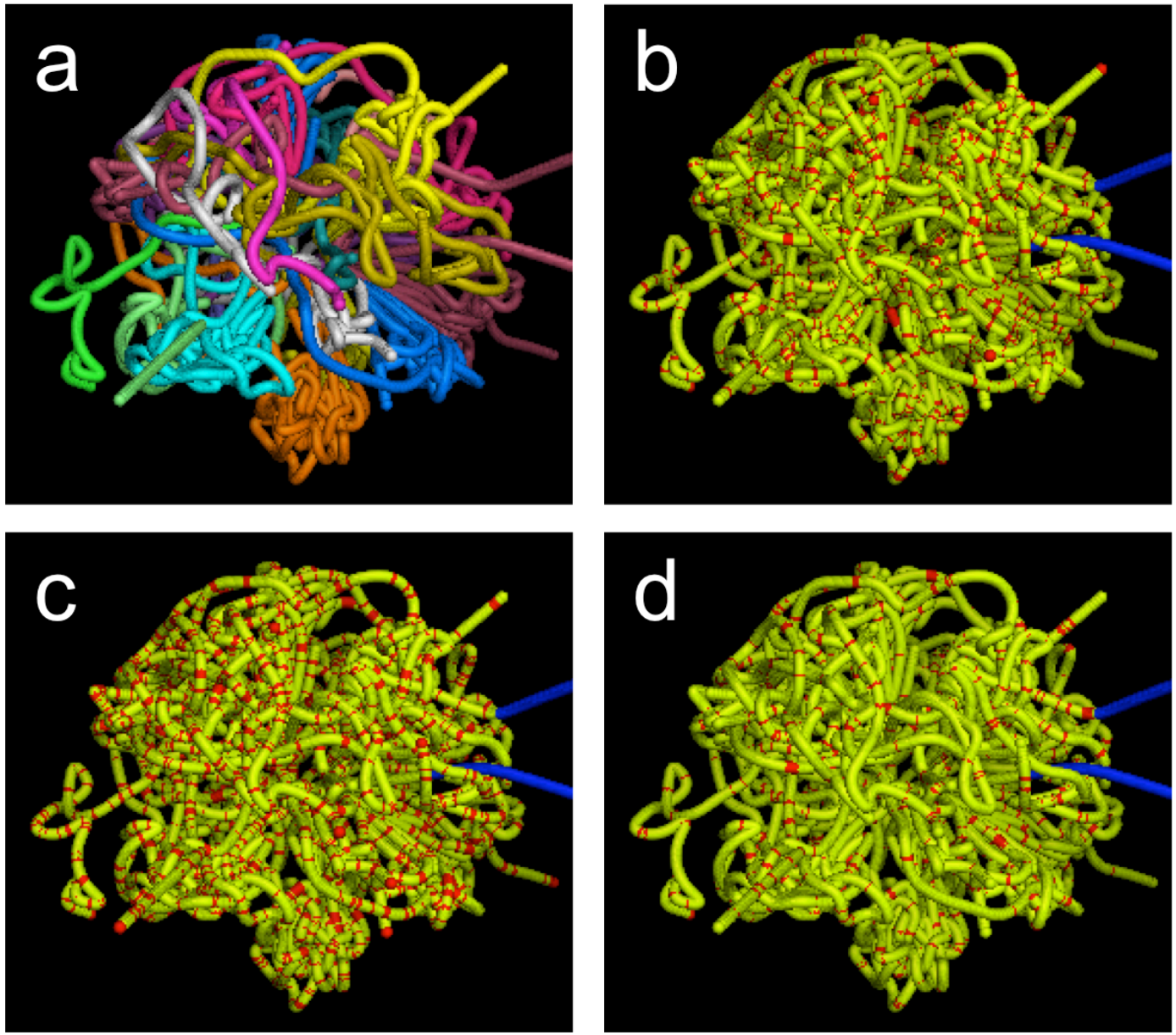


Diament, Tuller *PLoS CB* (2015)

Why 3D Reconstructions

- Improves on identifying co-localized functional elements *versus* contact maps Capurso, Segal *BMC Genomics* (2014)
  - multi-way *versus* pairwise
  - borrowed strength from contiguity
- Can readily superpose genome-indexed attributes -- problematic for contact pairs
- Find focal extrema: e.g. transcription factories, peripheral heterochromatic regions

Yeast with ChIP-Seq Overlays



Swi6

Pol2Ser5p

Tup1

Bump-Hunting

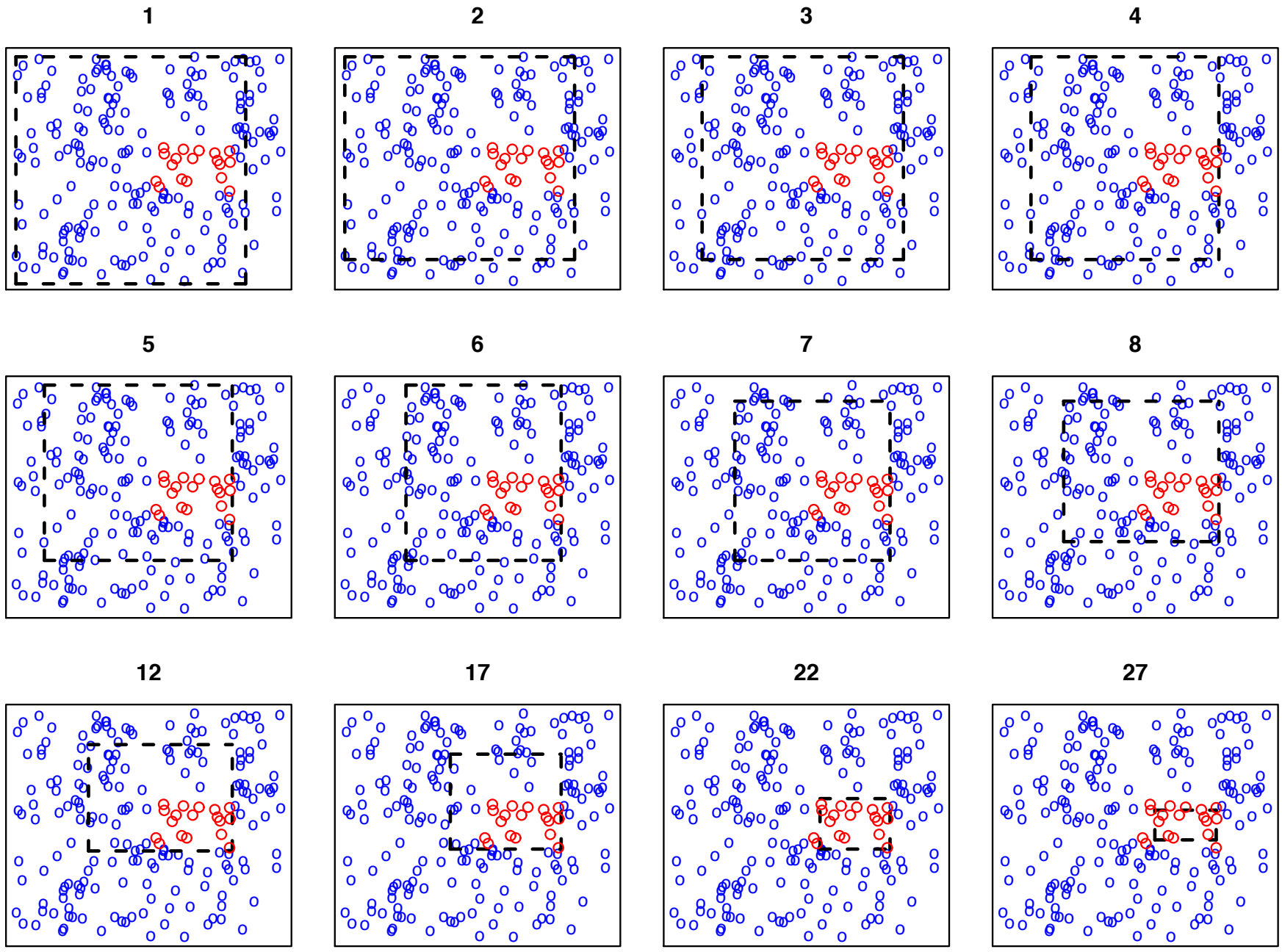
Function f; covariates $\mathbf{x}$. Goal: find covariate space subregions $\mathcal{R}$ *st* $\bar{f}_{\mathcal{R}} = \text{ave}_{\mathbf{x}\in\mathcal{R}} f(\mathbf{x}) \gg \bar{f}$

$\mathbf{x} = (x, y, z)$: coordinates; f: ChIP-Seq score

Want interpretable $\mathcal{R}$: impose $\mathcal{R} = \bigcup_{k=1}^{K} \mathcal{B}_k$; where each $\mathcal{B}_k$ is a "box": $\mathcal{B}_k = \bigotimes [a_{jk}, b_{jk}]$

Two-phase strategy used to find good boxes:
peeling – remove small unimportant regions;
pasting – enlarge boundaries of resultant box

Friedman, Fisher *Stat & Comp* (1999) R pkg: prim


1
2
3
4
5
6
7
8
12
17
22
27

ChIP-Seq: Yeast, swi6

Park et al *Plos One* (2013)

min_beads = 25

boxes = 670
sig = 10
sig & (min_chr > 0.2) = 6

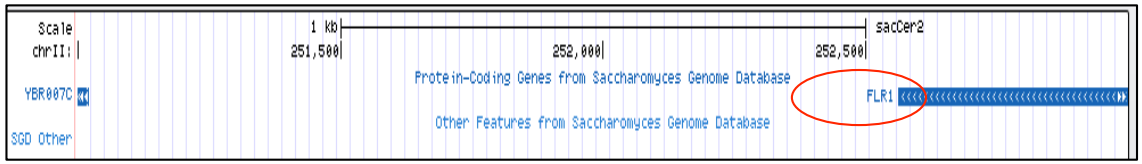
box_id	nbeads	mean_out	p_adj	min_chr
18	**45**	**0.514**	**6.71e-03**	**0.533**
160	25	0.894	6.71e-03	0.520
44	25	0.671	6.71e-03	0.440
60	25	0.785	6.71e-03	0.400
1	25	1.081	6.71e-03	0.080
87	25	0.665	1.33e-02	0.000
25	25	0.652	1.99e-02	0.560
42	25	0.651	1.99e-02	0.360
11	29	0.596	3.31e-02	0.000
125	26	0.645	4.63e-02	0.000

swi6_minbeads25_box18

3 regions from 3 chromosomes

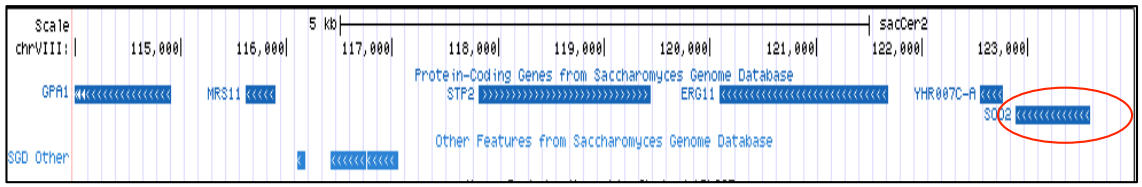
chrII:	251 kB – 252 kB	(3 beads)
chrVIII:	114 kB - 124 kB	(21 beads)
chrXIII:	259 kB – 270 kB	(21 beads)

chrII



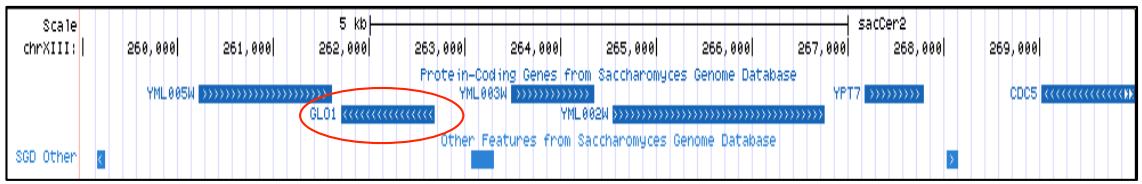
FLR1
fungicide transporter
(fungicide)

chrVIII



SOD2
superoxide dismutase
(reactive oxygen ROS)

chrXIII



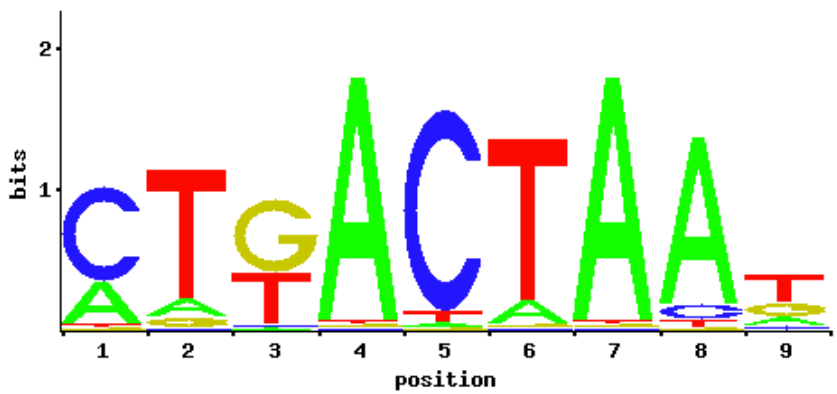
GLO1
glyoxylase
(methylglyoxal MG)

Downstream Analysis

- Each of the regions in box_18 contains a gene (FLR1, SOD2, GLO1) that becomes expressed in response to toxic compounds (fungicides, ROS, MG). The genes are:
  - Functionally similar
  - Repressed
  - Physically co-localized
- Potentially poised for co-activation
- Do they share a transcription factor??

Yap1 Motif Finding in box18 regions

chrII:	251 kB – 252 kB	(3 beads)
chrVIII:	114 kB - 124 kB	(21 beads)
chrXIII:	259 kB – 270 kB	(21 beads)

TF	Logo
YAP1_D Matrices: PFM PCM PWM Data Source: Foat et. al.	foat.YAP1_D binding site logo 

Position Frequency Matrix for Foat et. al. YAP1_D:

A	0.30	0.11	0.01	0.97	0.02	0.12	0.97	0.88	0.15
C	0.69	0.01	0.01	0.01	0.92	0.01	0.01	0.08	0.07
G	0.01	0.06	0.58	0.01	0.01	0.01	0.01	0.01	0.25
T	0.01	0.82	0.40	0.01	0.04	0.87	0.01	0.03	0.53

The MEME Suite

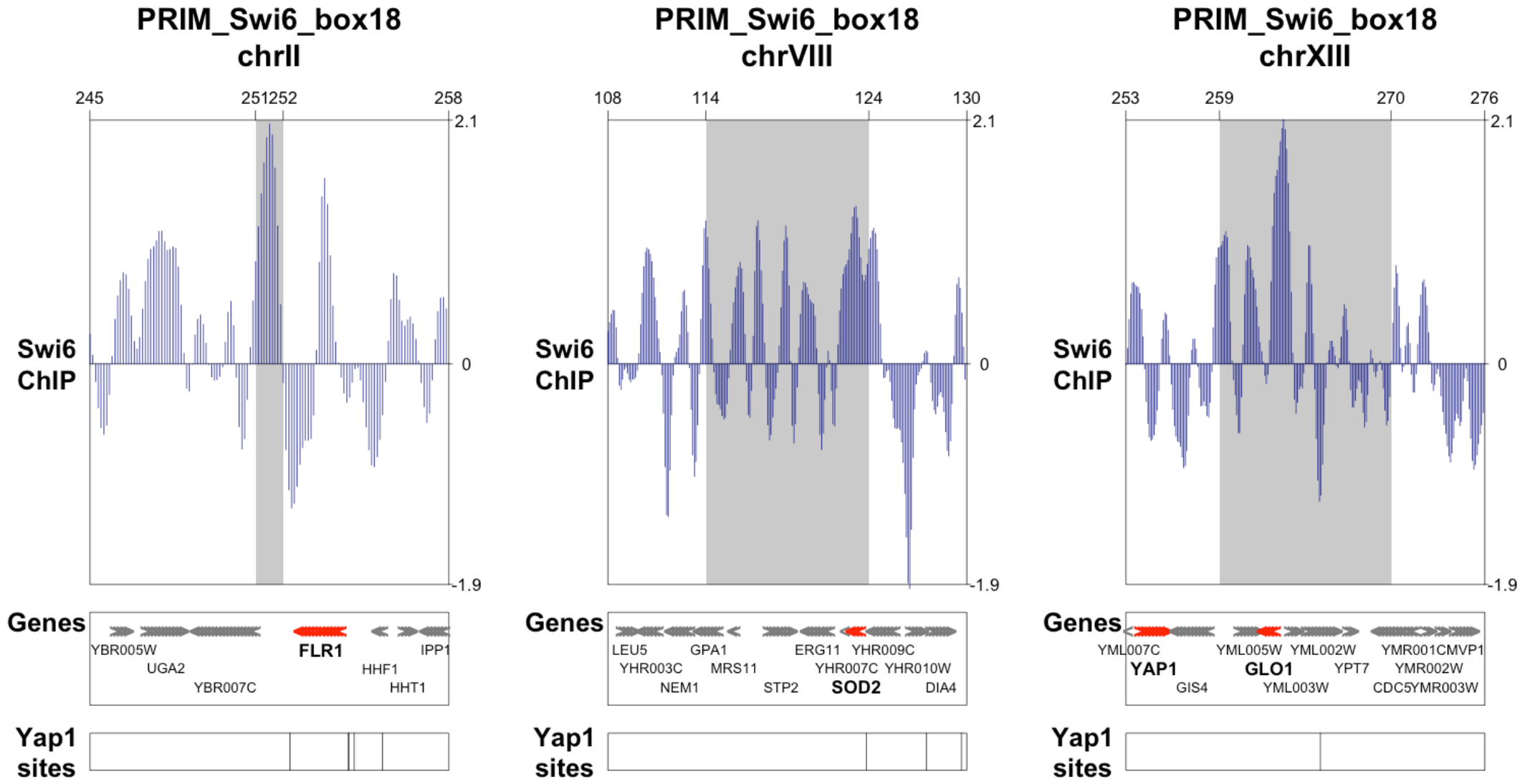
Motif-based sequence analysis tools

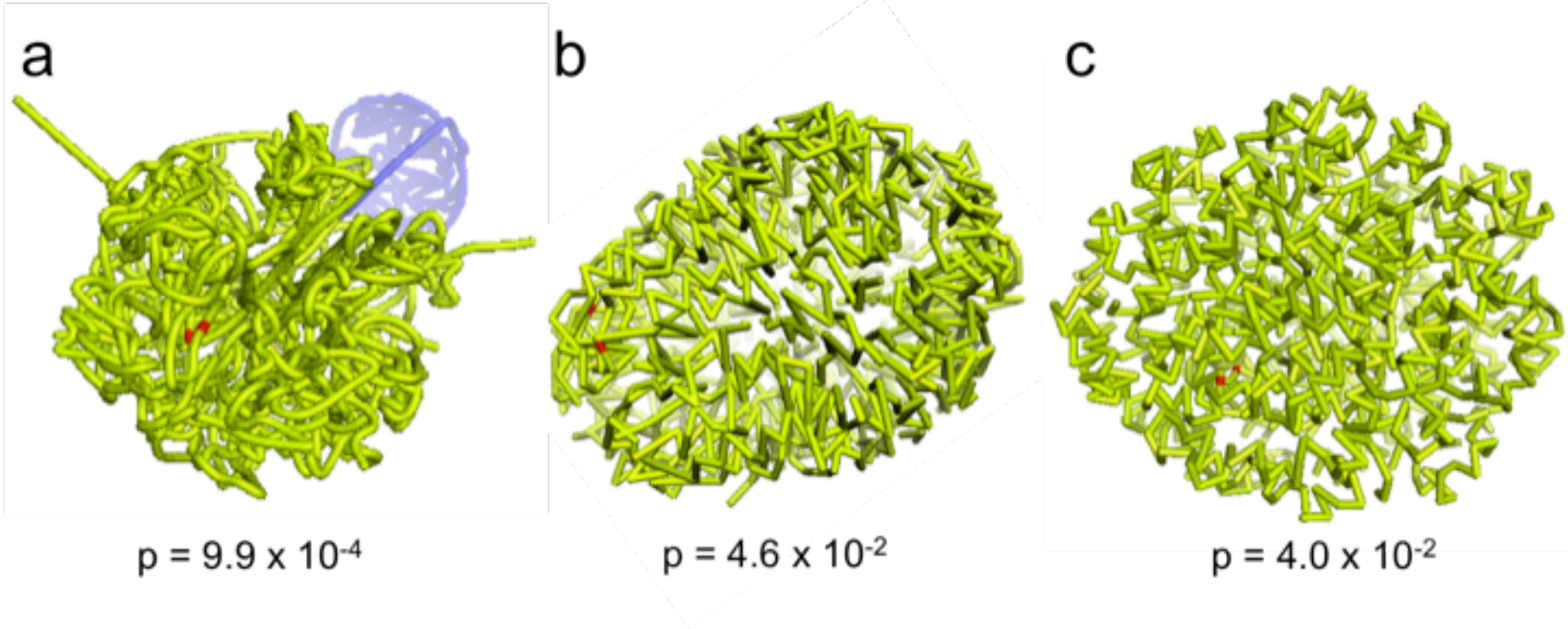
FIMO

Find Individual Motif Occurences

Downstream Analysis

- FLR1, SOD2, GLO1 are activated by the same transcription factor Yap1
- "The *S. cerevisiae* transcription factor Yap1 plays an important role in oxidative stress response and multidrug resistance by activating target genes involved in cellular detoxification."
- Nguyen et al *J Biol Chem* (2001)


PRIM_Swi6_box18
chrII
245
251252
258
2.1
0
-1.9
Swi6
ChIP
Genes
YBR005W
UGA2
YBR007C
FLR1
HHF1
HHT1
IPP1
Yap1
sites
PRIM_Swi6_box18
chrVIII
108
114
124
130
2.1
0
-1.9
Swi6
ChIP
Genes
LEU5
YHR003C
NEM1
GPA1
MRS11
STP2
ERG11
YHR007C
SOD2
YHR009C
YHR010W
DIA4
Yap1
sites
PRIM_Swi6_box18
chrXIII
253
259
270
276
2.1
0
-1.9
Swi6
ChIP
Genes
YML007C
YAP1
GIS4
YML005W
GLO1
YML003W
YML002W
YPT7
YMR001CMVP1
YMR002W
CDC5YMR003W
Yap1
sites



Duan	Varoquaux	Varoquaux
Constrained MDS	MDS	MDS + Poisson
Explicit Factor	Matrix Balance	Matrix Balance

Reconstruction-free Hotspots

- Considerable uncertainty still surrounds inferred 3D genome reconstructions
- Developing methods to elicit hotspots without requiring a reconstruction desirable
- Problematic since hotspots are critically dependent on 3D proximity

Reconstruction-free Hotspots

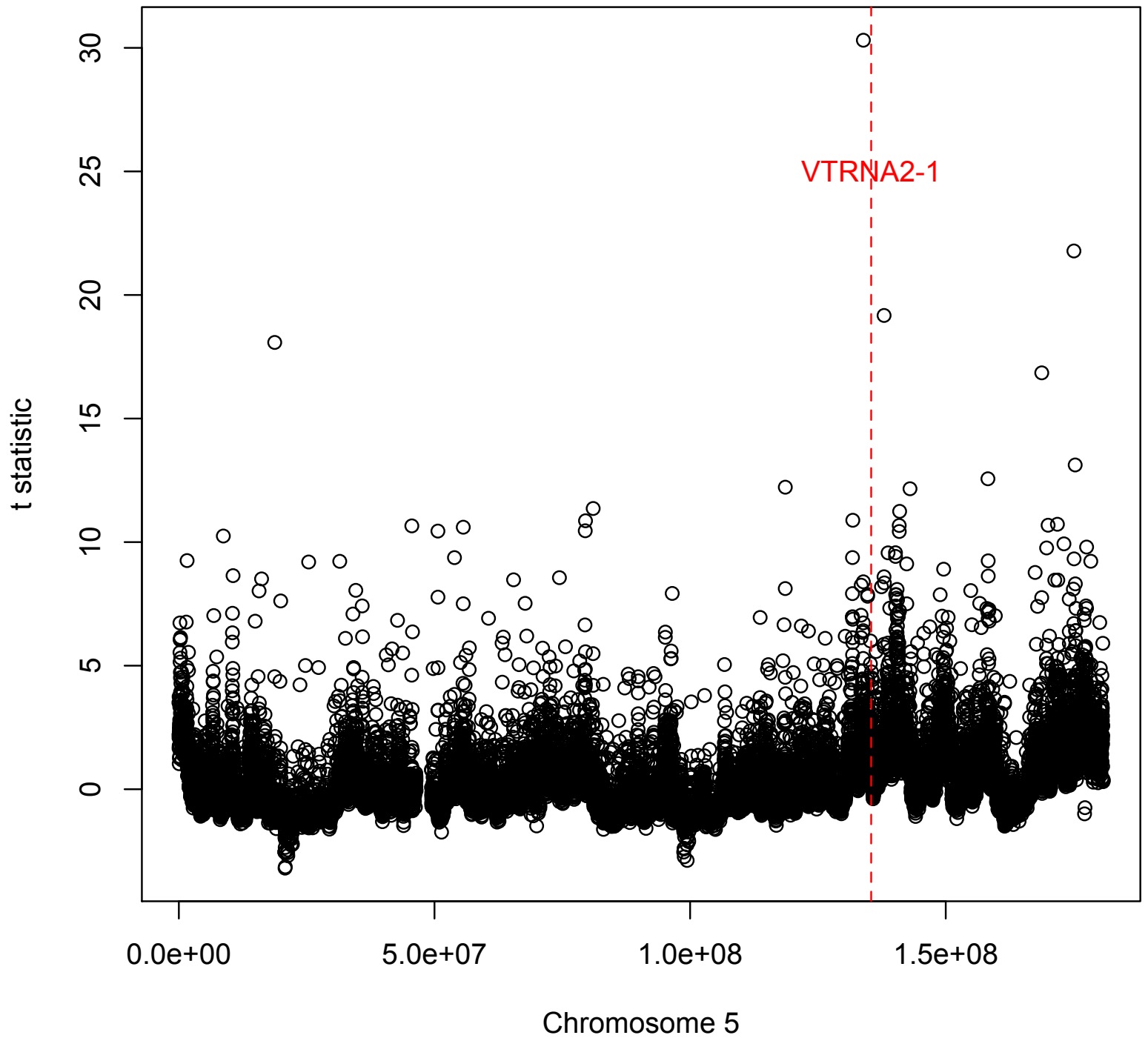
Distribute response Y according to contacts F:

$\tilde{Y}_i = \sum_{j \in \mathcal{S}} g(F_{ij}, Y_j) \to \sum_{j \in \mathcal{S}} F_{ij} \cdot Y_j$

Large $F \implies$ small D: proximal upweighting

Further control – mimic $\mathcal{B}_k$ – through refining $\mathcal{S}$

Rank $\tilde{Y}_i$s; inference via permutation


VTRNA2-1
t statistic
0
5
10
15
20
25
30
0.0e+00
5.0e+07
1.0e+08
1.5e+08
Chromosome 5

Future Work

- Refining, tuning, accelerating MDS, others
- Sampling strategies for two-stage algorithm:
  - Bi-clustering to optimize inter-chromosomal information
- Evaluating reconstruction accuracy and reproducibility:
  - Multi-chromosome, multi-plex FISH
  - Generating null referent distributions

Future Work

- Rotation invariant response analyses:
  - tuning nearest neighbor methods
  - recursive partitioning with hyperplanes
  - persistence homology: Betti numbers, barcodes of excursion sets
- Methodology for reconstruction-free hotspots
- Design, analysis and reconstruction for single cell and in situ assays:
  - replicates, perturbations, time-course