A Principal Curve Approach to Three-Dimensional Chromatin Configuration Reconstruction

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

Center for Bioinformatics & Molecular Biostatistics

BIRS 2018, Oaxaca

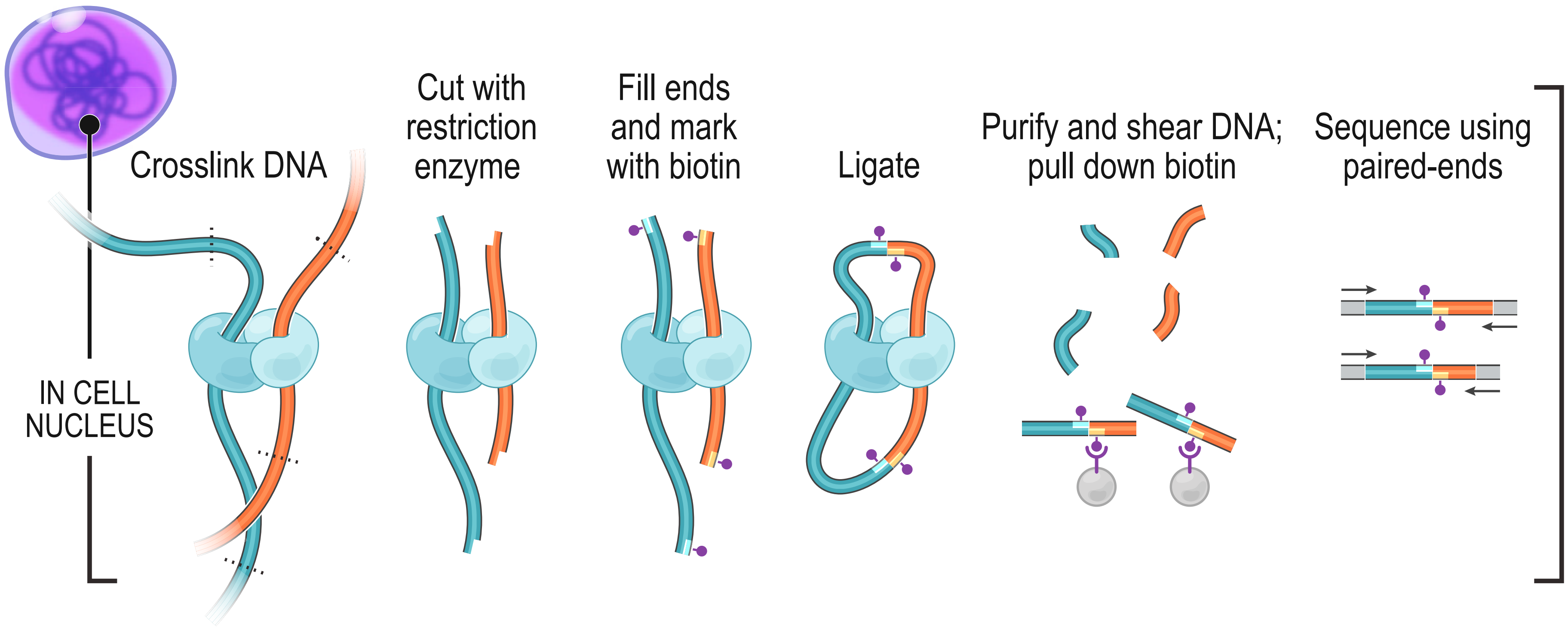
Importance of 3D Conformation

- 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

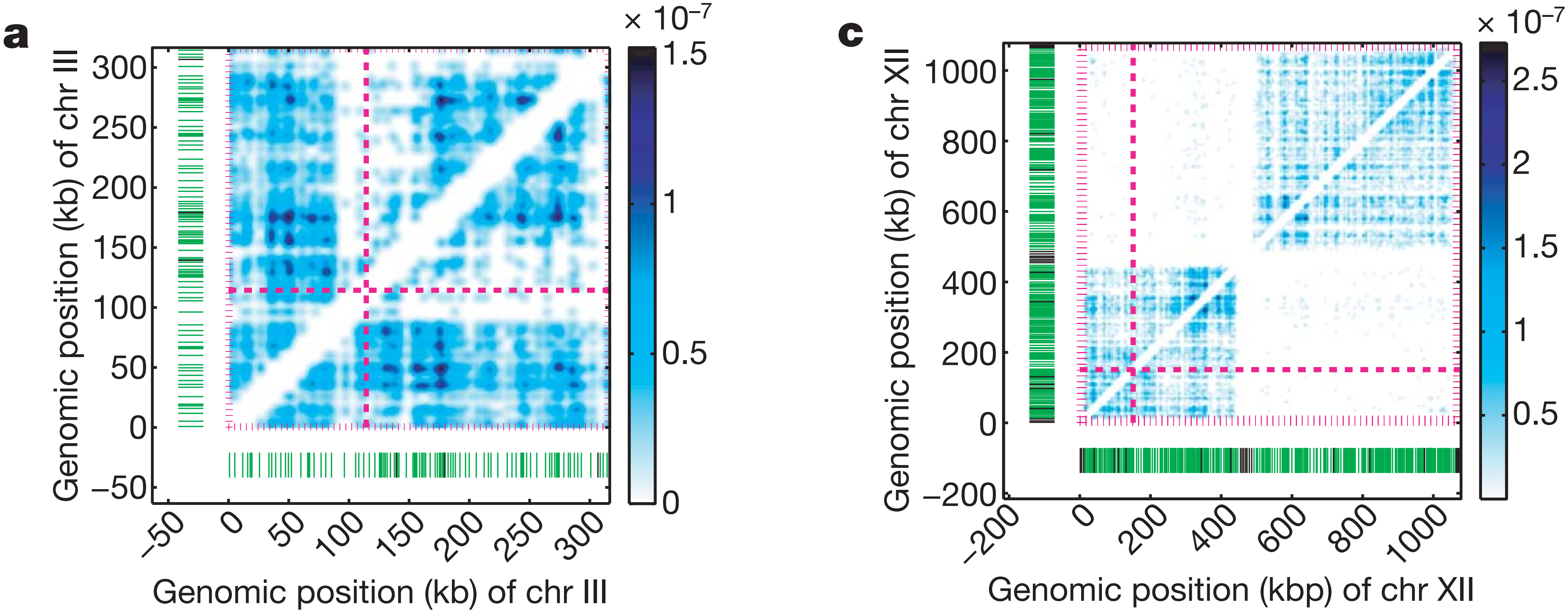
- Challenging at even modest resolutions:
    - genomes are highly condensed
    - genomes are dynamic, variable
    - traditional assays are low throughput and low resolution (FISH coarse)
- Recently devised suite of Chromatin Conformation Capture techniques has

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


Crosslink DNA
Cut with restriction enzyme
Fill ends and mark with biotin
Ligate
Purify and shear DNA; pull down biotin
Sequence using paired-ends
IN CELL NUCLEUS

Performed using large – $\sim 10^6$ – cell populations

Output: Contact / Interaction Maps

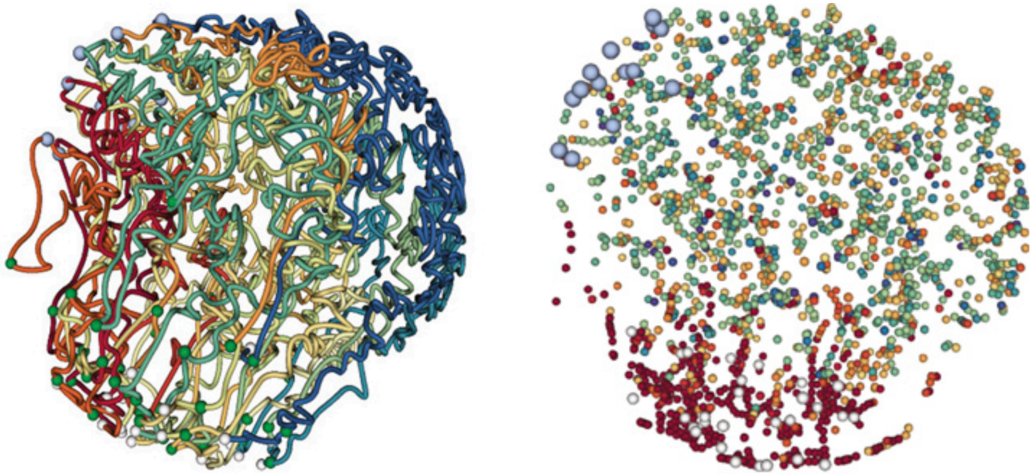


Also *inter*-chromosomal maps.
Resolution determined by binning.

From Contacts to 3D Structure

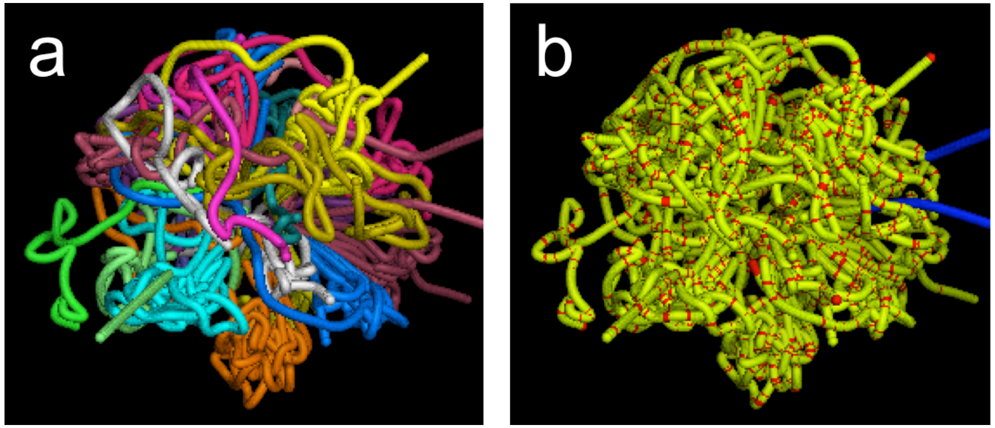
- Objective: given contact matrix *C*, obtain a 3D structure (or an ensemble thereof) the between-loci pairwise distances of which recapitulate corresponding contact counts.
- Many reconstruction algorithms advanced.
- Despite assumptions, uncertainties added value derives from inferring 3D architecture:
    - In part derives from super-posing genome attributes on the reconstruction.

3D *P. falciparum* : Overlaid Expression



Can demonstrate expression-telomere distance gradient. But what about detecting focal regions : 3D hotspots ?

3D *S. cerevisiae* : Overlaid ChIP-Seq

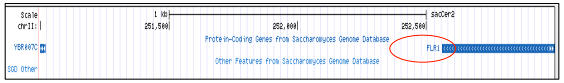


swi6_minbeads25_box18

3 regions from 3 chromosomes

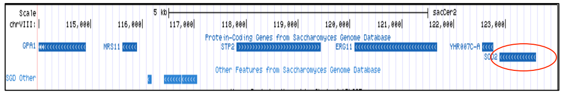
chrII:	251 kB – 252 kB	(3 beads)
chrVIII:	114 kB - 124 kB	(21 beads)
chrXIII:	259 kB – 270 kB	(21 beads)

chrII


Scale
1 kb
sacCer2
chrII:
251,500
252,000
252,500
Protein-Coding Genes from Saccharomyces Genome Database
YBR007C
FLR1
Other Features from Saccharomyces Genome Database
SGD Other

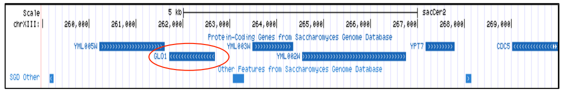
FLR1
fungicide transporter
(fungicide)

chrVIII



SOD2
superoxide dismutase
(reactive oxygen ROS)

chrXIII



GLO1
glyoxylase
(methylglyoxal MG)

Optimization / Consensus Methods

- Generally utilize two steps:
    - convert *C* into a distance matrix *D* that captures expected pairwise distances
        - differing assumptions; $D \propto C^{-\alpha}; \quad \alpha > 0$
        - sometimes interplay with second step
    - learn / estimate 3D structure from *D*
        - multi-dimensional scaling (MDS) criteria
        - weights, non-metric, constraints, ...
        - algorithms include SA, IPO, SDE, MM...

Distances to 3D Structure

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

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

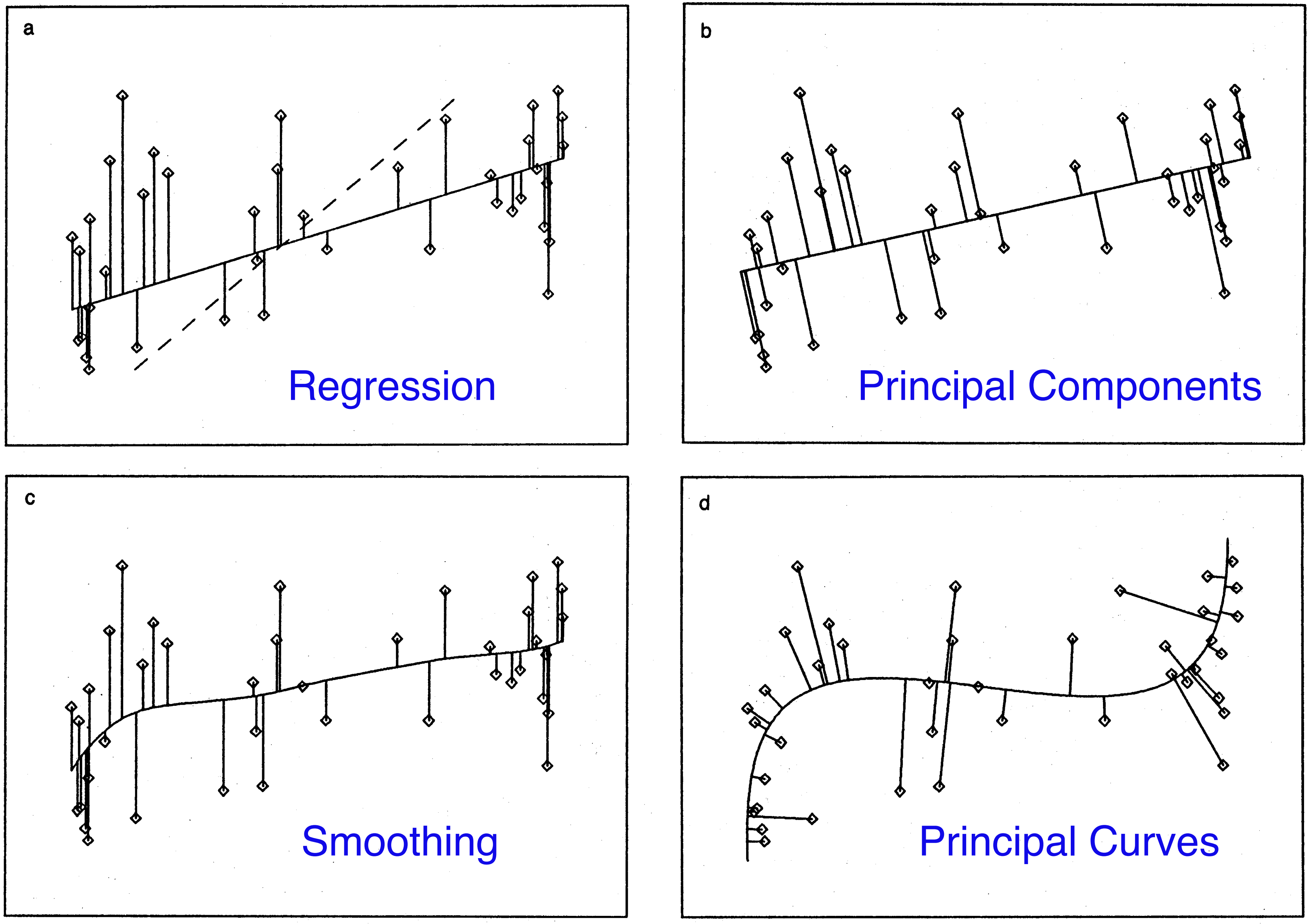
- Penalty: $\tau \sum_{\{i,j \mid D_{ij} = \infty\}} \|x_i - x_j\|^2$

  - Non-interacting loci cannot be too close

Constraints and Contiguity

- Many biological constraints can be imposed:
    - Yeast: centromere clustering, 1um sphere.
- Constraints are difficult to specify; cell-type, resolution specific; increase compute burden.
- Malaria: adjacent 10kb loci within 91nm; Yeast: adjacent 10kb loci > 30nm.
    - Indirect way of imposing contiguity.
- Here we directly prescribe that the solution, per chromosome, is a 1D smooth curve.

Principal Curves



Regression

Principal Components

Smoothing

Principal Curves

Principal Curve Metric Scaling

Goal: 1D curve f in R^3 with inner products between n points on f approximating $C_{n\times n}$.

$f(\lambda)$: vector fn with 3 components; λ 1D index.

Genomic coordinates

Want coordinate functions to be smooth wrt λ so we represent each using a spline basis:

$$f_{ij}(\lambda) = \sum_{k=1}^{K} h_{ik}(\lambda)\theta_{kj},\ j = 1,2,3;\ i = 1,\ldots,n$$

where K is the number of knots $\sim$ spline df.

$F = H\Theta$ where Θ is $K \times 3$ matrix of coefficients.

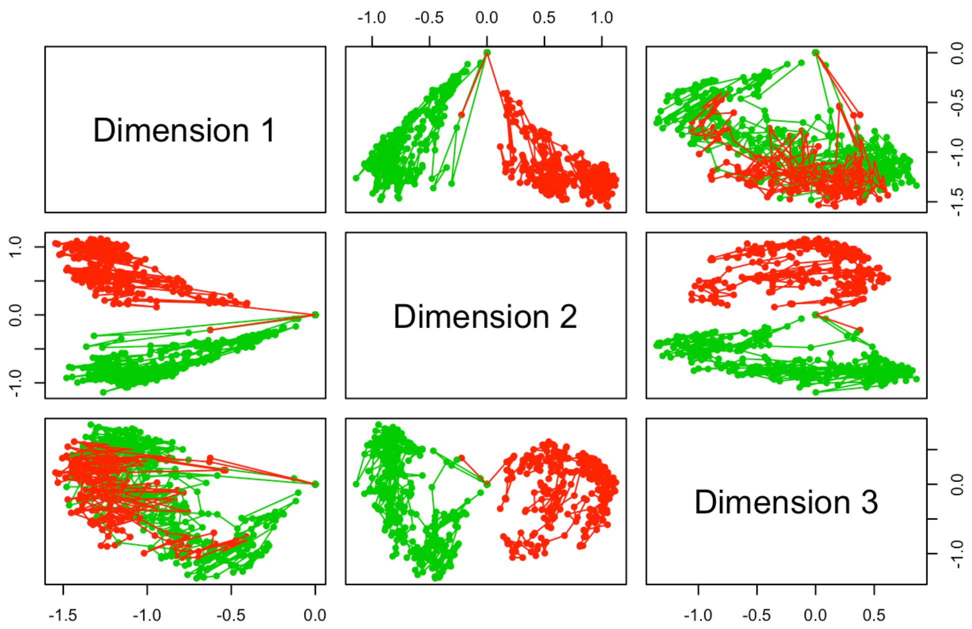
WLOG assume H is orthonormal.

Metric scaling problem: $\min_\Theta \|C - H\Theta\Theta^T H^T\|_F^2$.

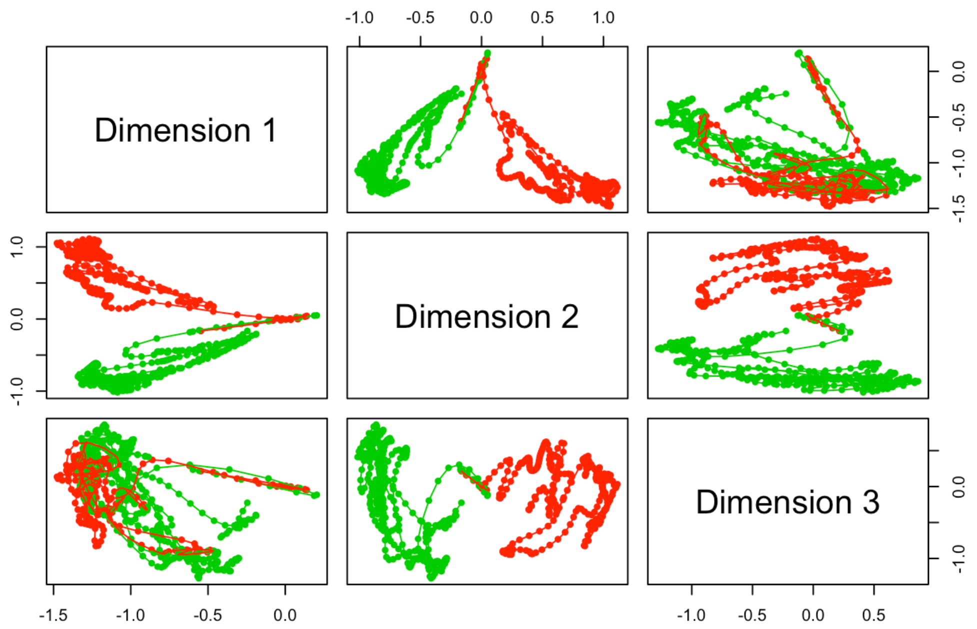
This is equivalent to $\min_\Theta \|H^T C H - \Theta\Theta^T\|_F^2$

which is solved by eigen-decomposition of $H^T C H$.

Df = 625 R-squared = 0.78

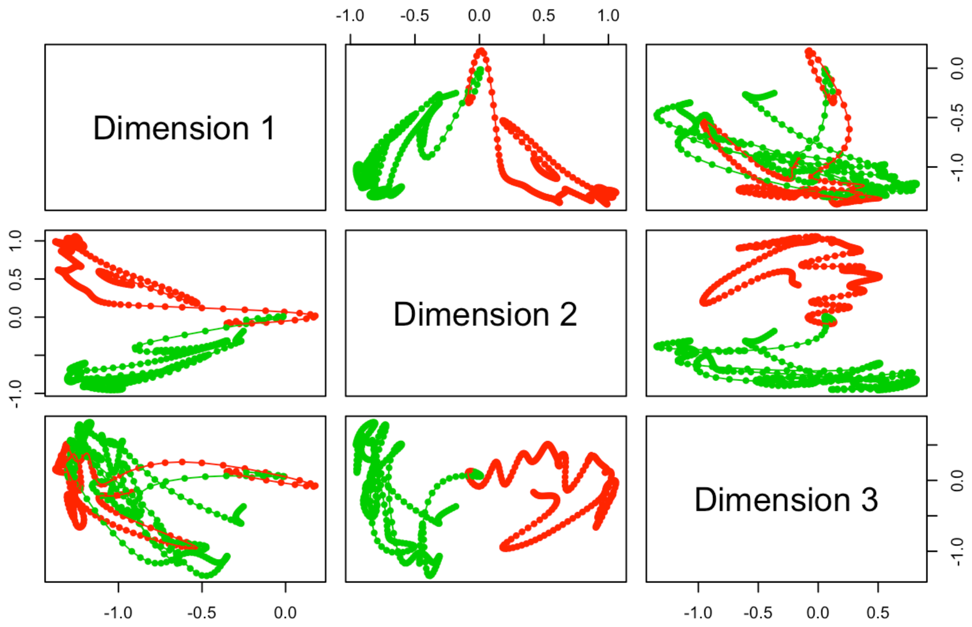


Df = 150 R-squared = 0.76

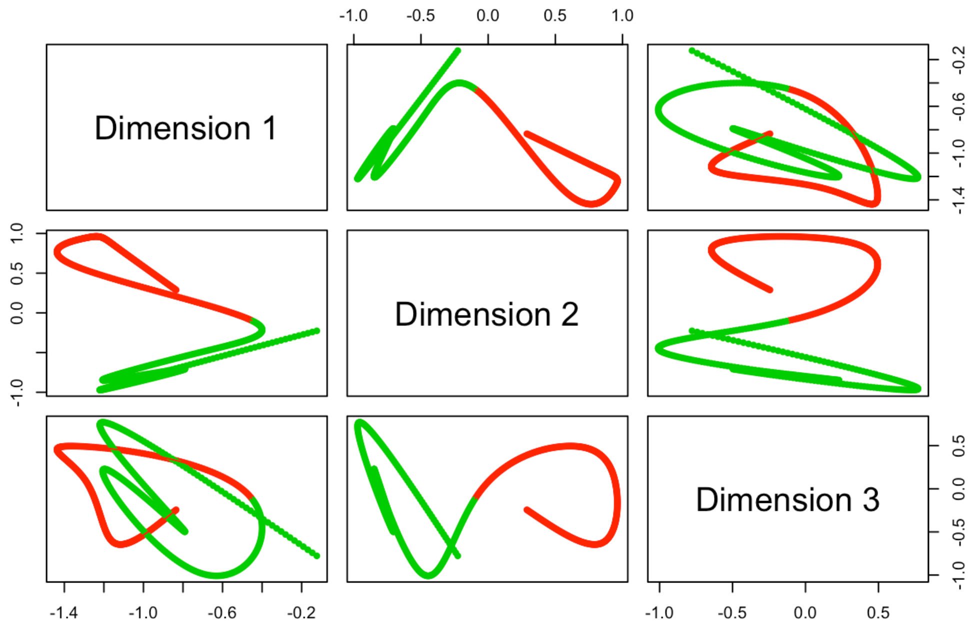


IMR90 // Chromosome 20 //100kb // Primary Series

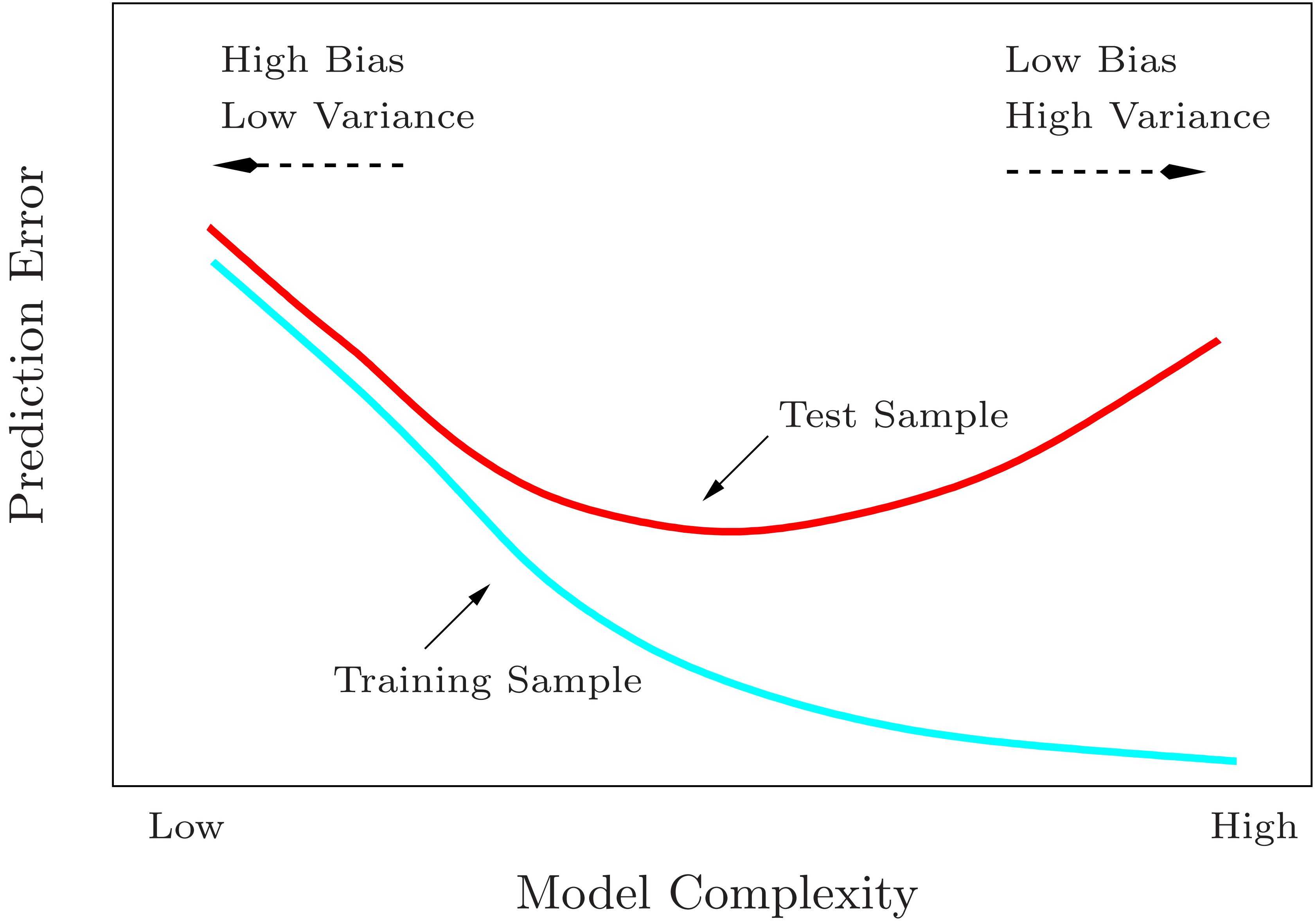
Df = 75 R-squared = 0.75



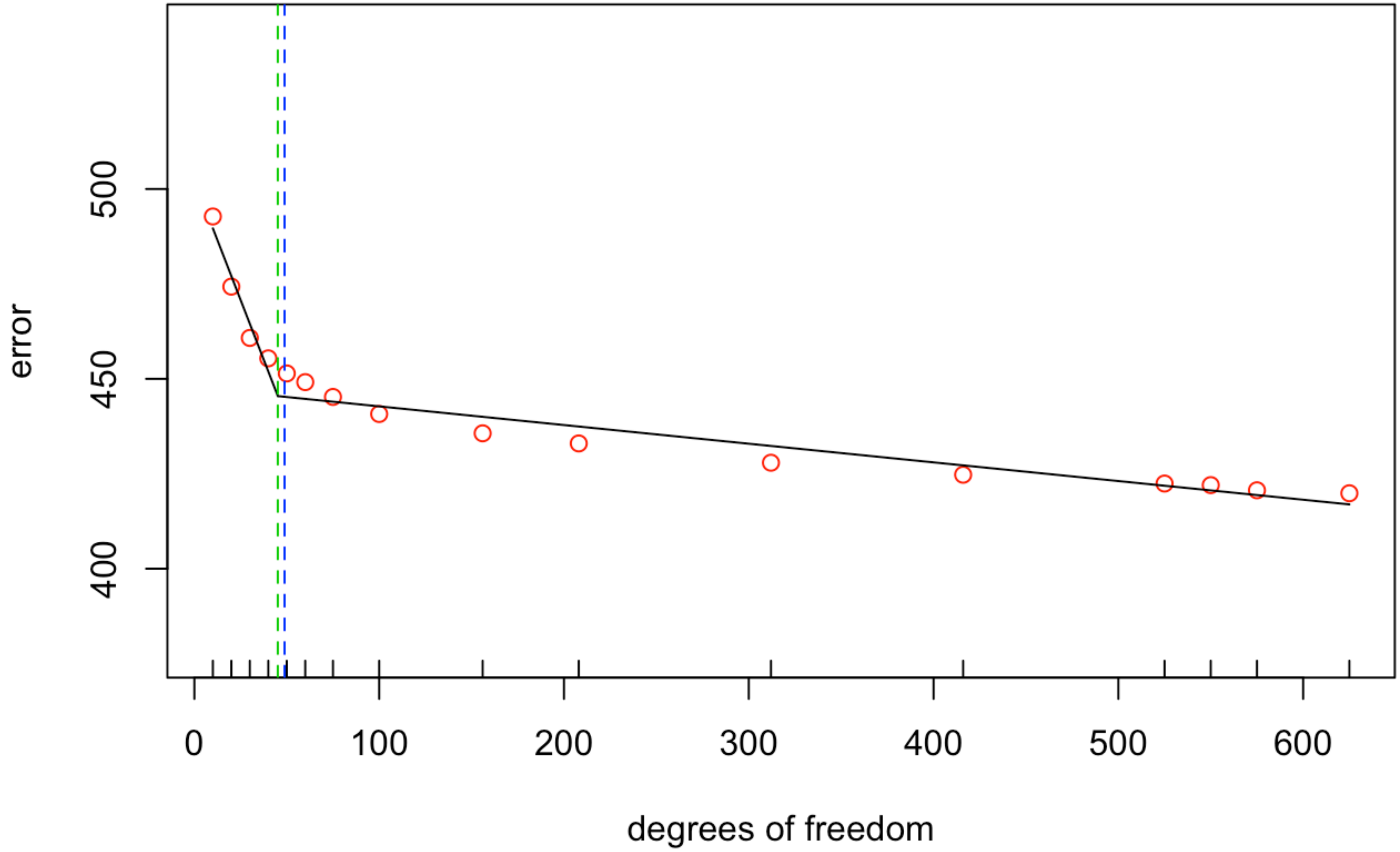
Df = 10 R-squared = 0.69



Determining Degrees-of-Freedom



Determining Degrees-of-Freedom



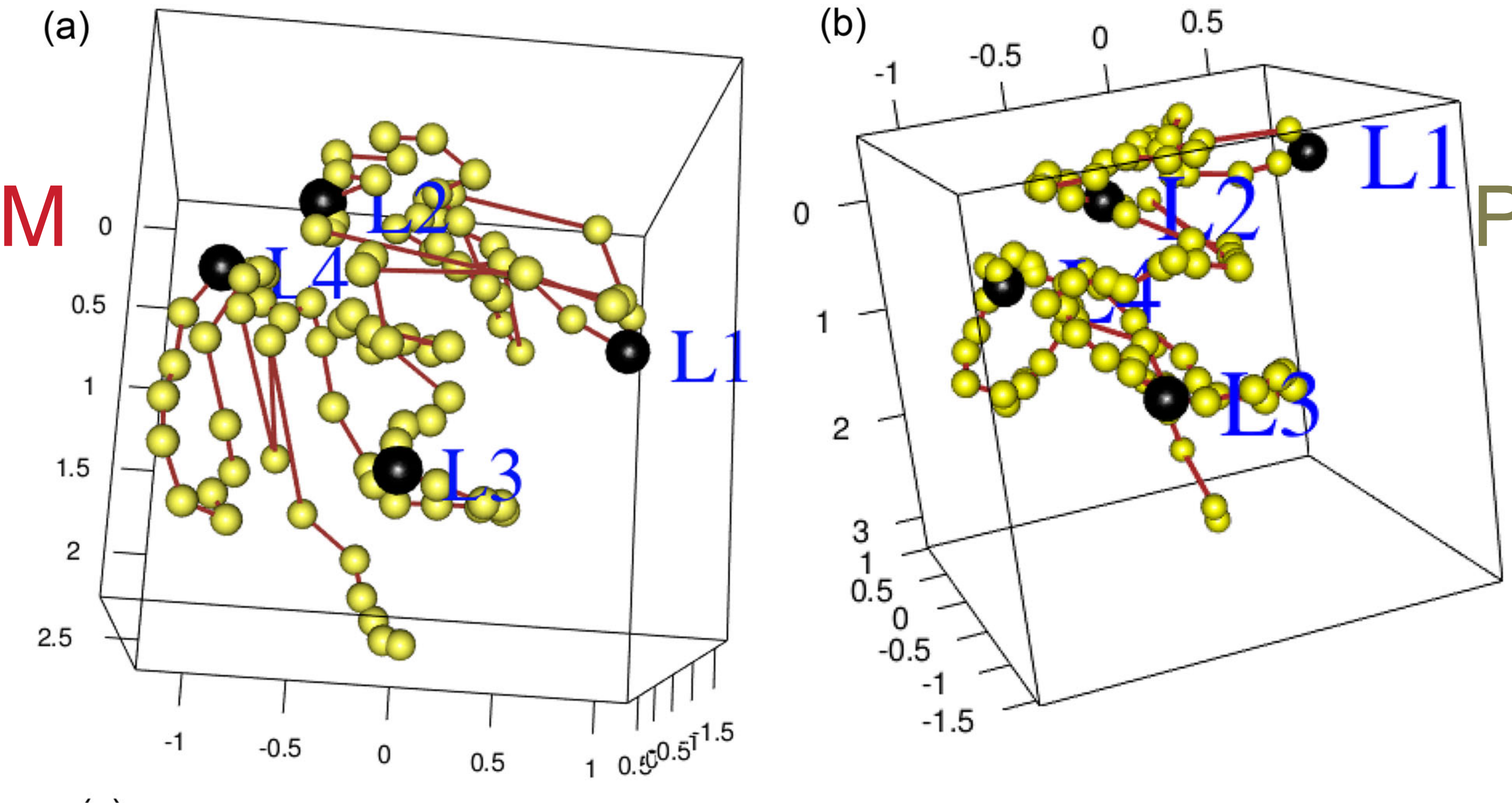
Broken-line / segmented regression: knot / elbow identification

Assessing Reconstruction Accuracy

- Challenging in view of absence of gold standards
    - reproducibility assessment based on replicates from differing RE digests
- Use of FISH: compare inter-probe distances
    - exceedingly limited due to probe sparsity
- Multiplexed FISH affords new possibilities

Standard FISH: 1Mb Resolution

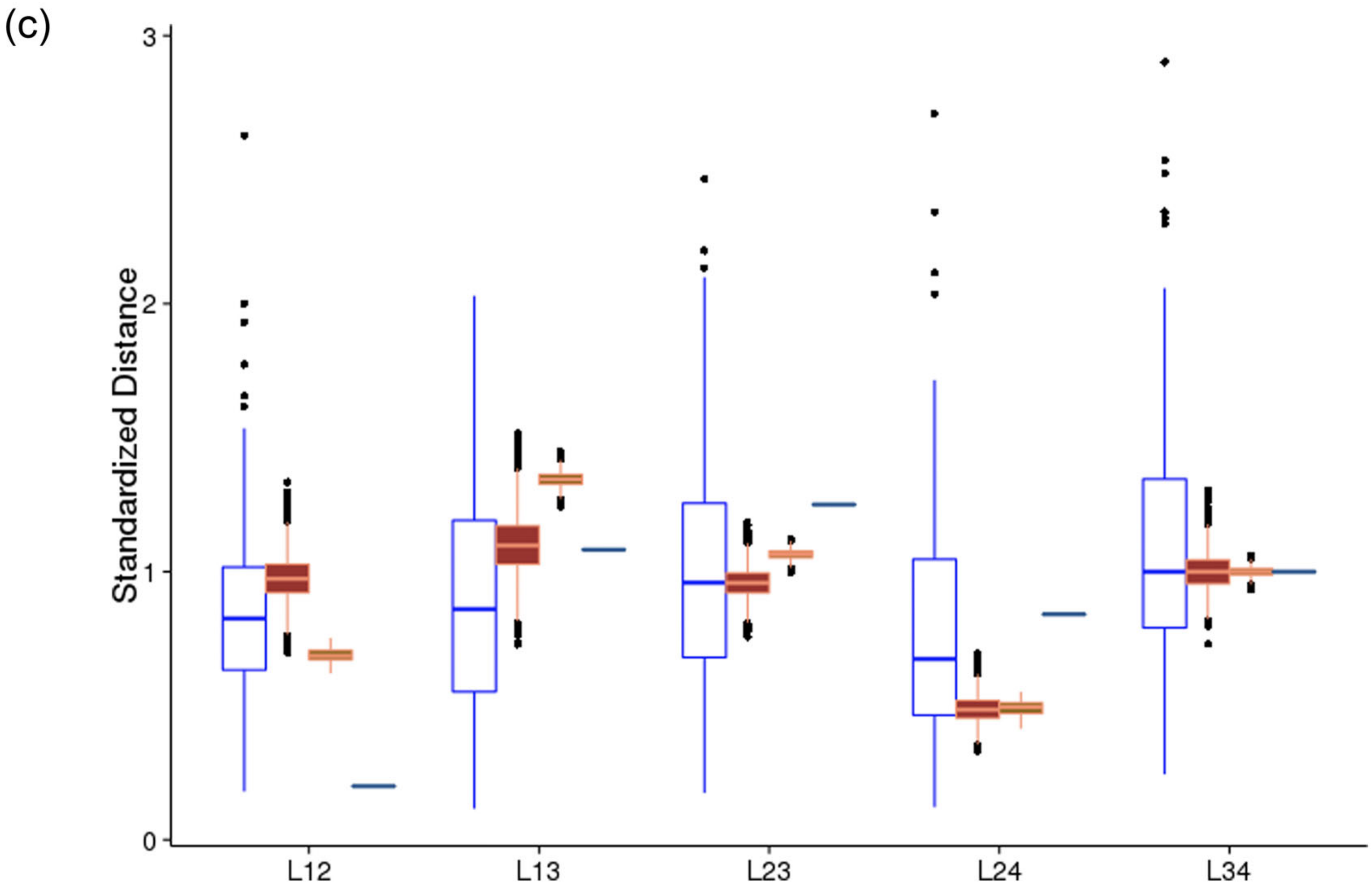
PRAM



PAM

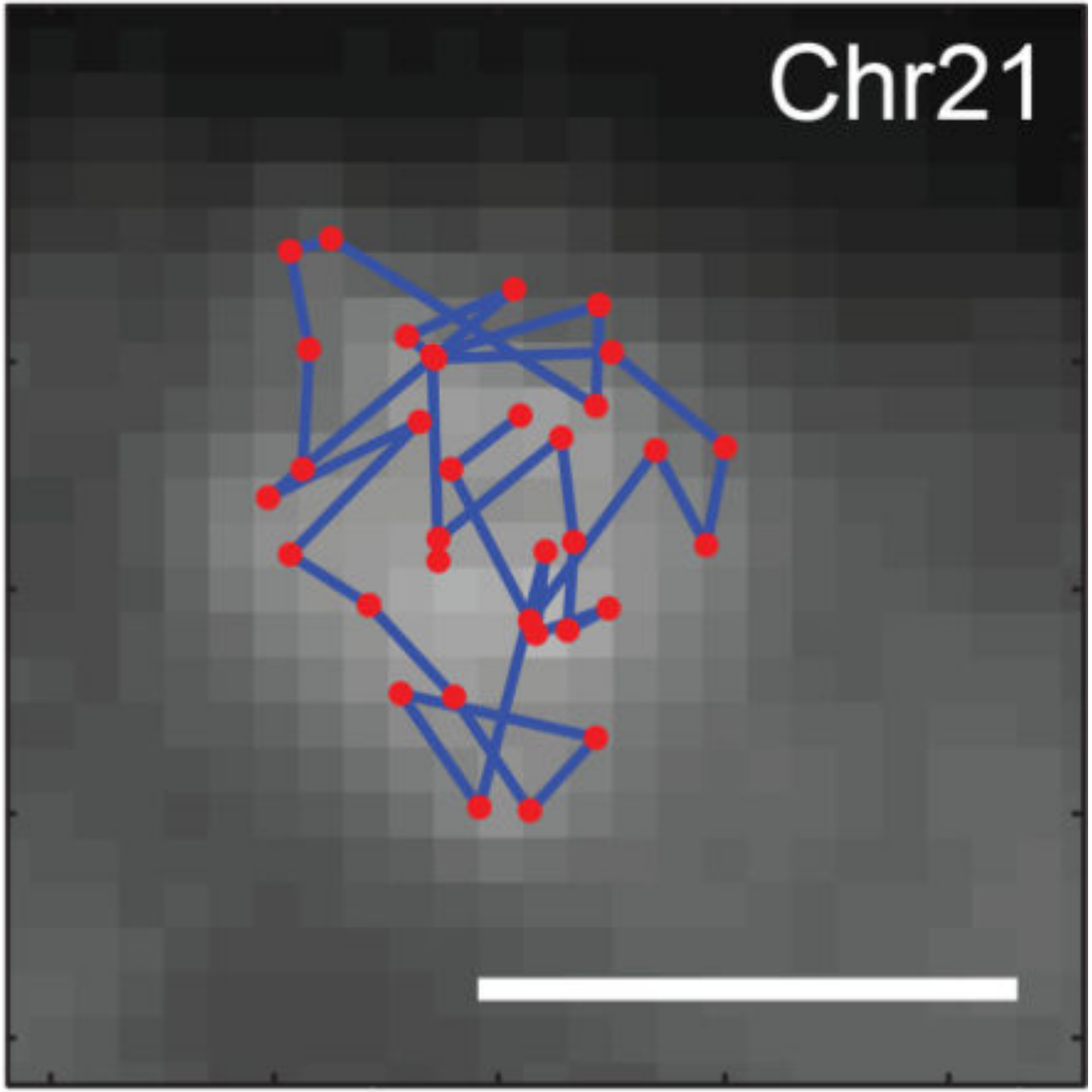
Park, Lin *Biometrics* (2016)

Standard FISH: 1Mb Resolution



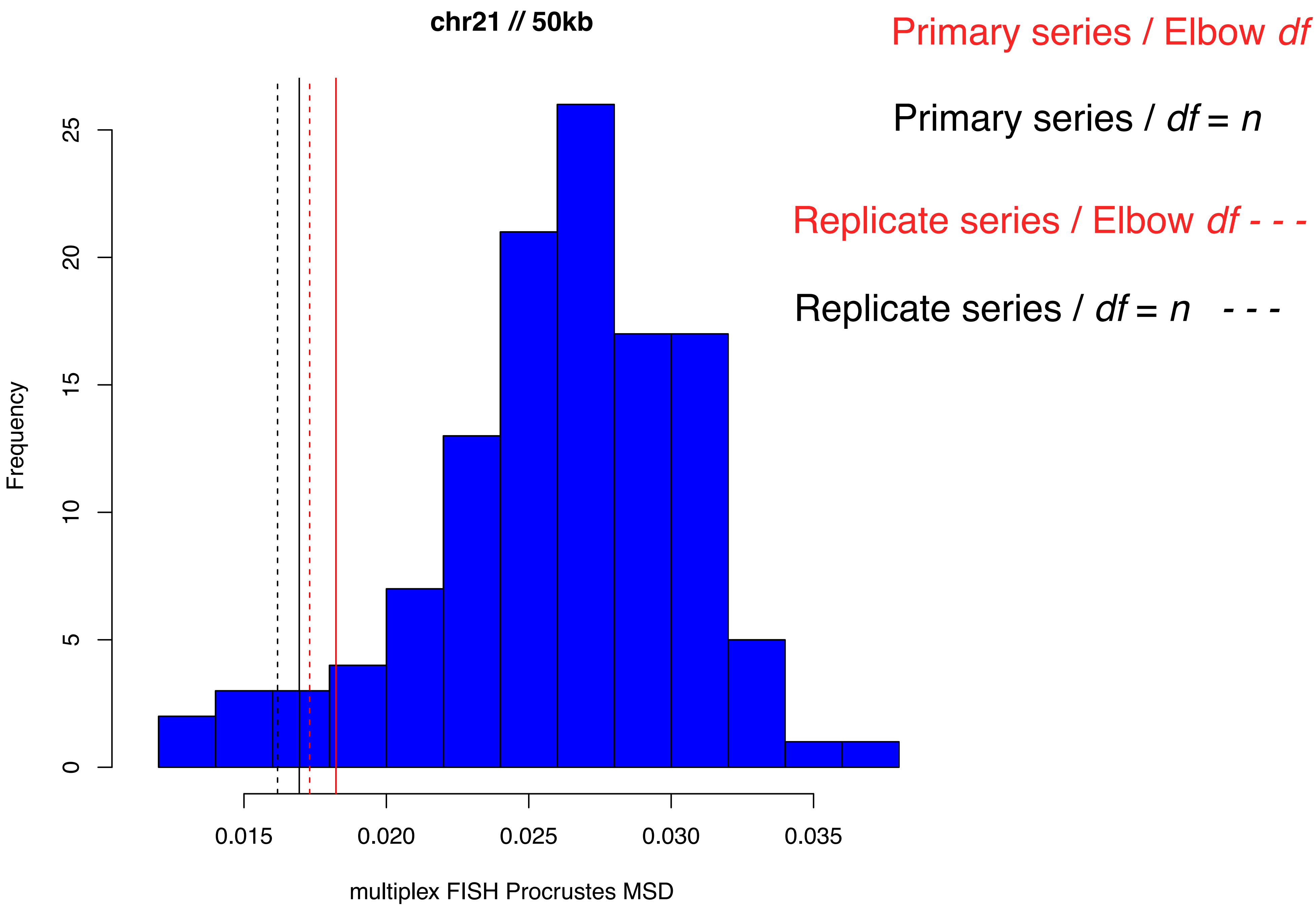
FISH, PRAM, PAM, ShRec3D

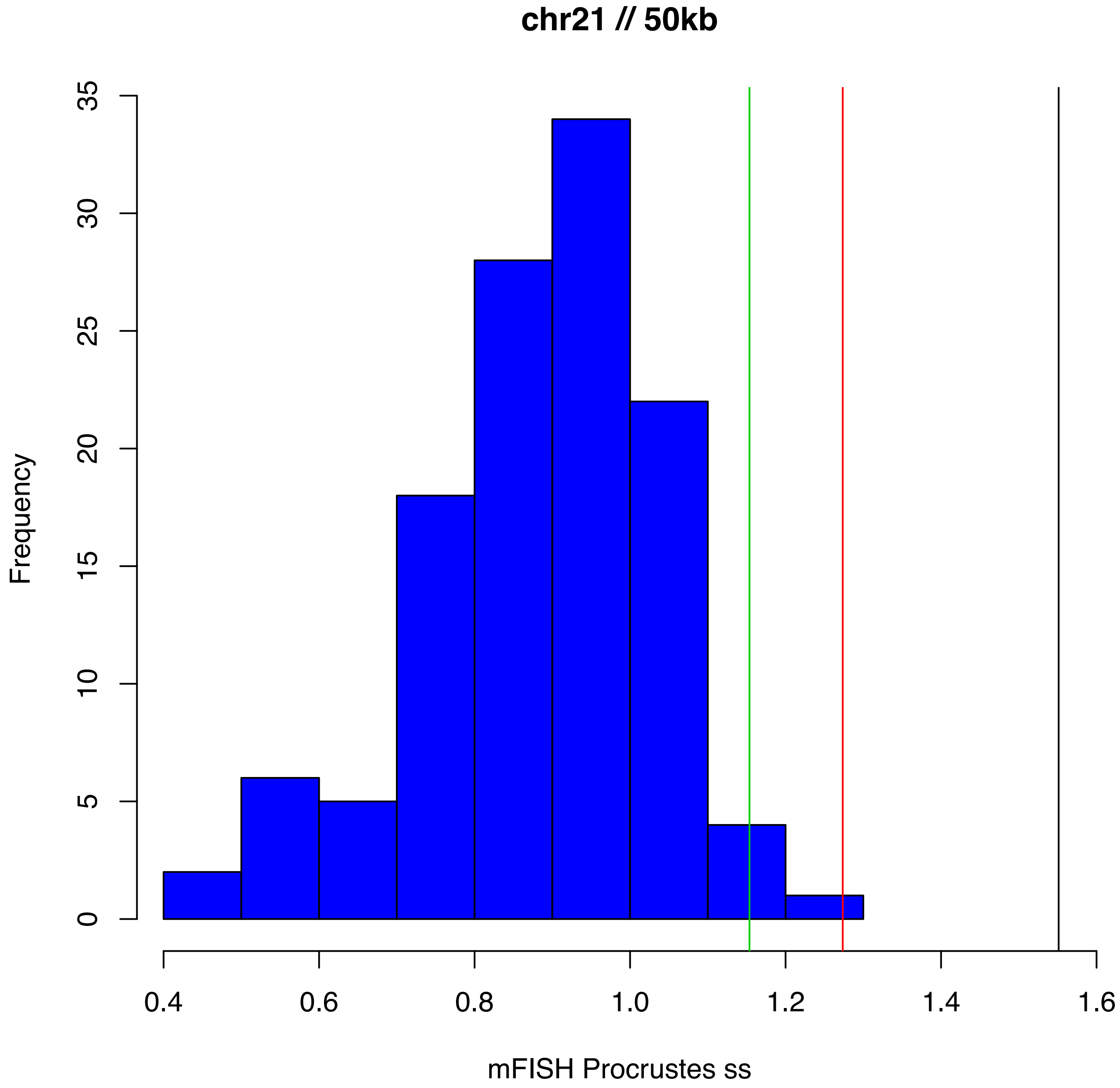
Multiplex FISH: 100kb Resolution



Multiplex FISH Assessments

- Crucial is existence of numerous replicates
  - provides natural referent distribution of (R)MSD distances
  - necessary in absence of thresholds (as per protein folding) or theoretic models
- For IMR90 cells have 111, 120, and 151 replicates for chromosomes 20, 21 and 22.
- Here evaluate 3D reconstruction obtained via PCMS algorithm using IMR90 Hi-C data.


chr21 // 50kb
Primary series / Elbow *df*
Primary series / *df* = *n*
Replicate series / Elbow *df* - - -
Replicate series / *df* = *n* - - -
Frequency
0
5
10
15
20
25
0.015
0.020
0.025
0.030
0.035
multiplex FISH Procrustes MSD



Alternate algorithm — HSA: primary, replicate, combined

Future Work

- Degrees-of-freedom via cross-validation.
- Alternate bases (e.g. wavelets) or partitioning methods to capture hierarchical chromatin organization.
- Alternate transformations of *C*.
- Single-cell Hi-C.

Acknowledgements

- Trevor Hastie
- Elena Tuzhilina

- R01GM109457