Topological domains in chromatin

Carl Kingsford

Carnegie Mellon University

Joint work with Darya Filippova, Rob Patro, Geet Duggal, Emre Sefer, Brad Solomon

Thanks

Darya Filippova

Rob Patro

Geet Duggal

Emre Sefer

Funding

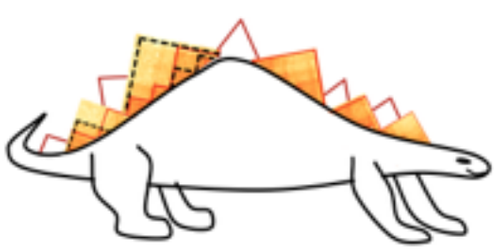
NIH R01 HG007104, R21 HG006913, T32 EB009403

NSF CCF-1256087, CCF-1319998

Sloan Research Fellow (C.K.)

Gordon and Betty Moore Foundation - Data Driven Discovery Investigator

Brad Solomon

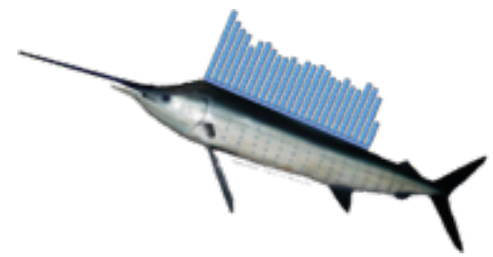


www.cs.cmu.edu/~ckingsf/software/armatus

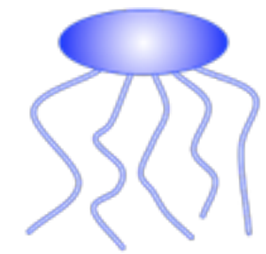
Our Recent Open-Source Work on Large-Scale Genomics

Identifying topological domains in Hi-C

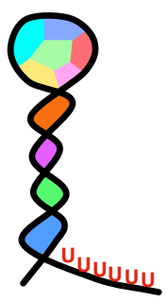
Finding confident structures in Hi-C



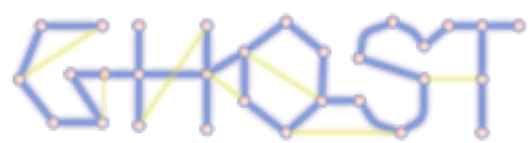
Measuring gene expression



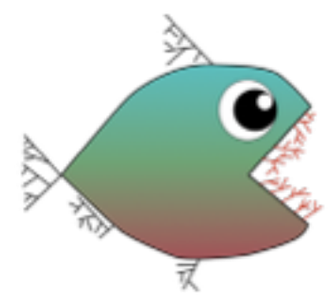
Counting kmers (part of Celera & Trinity Assemblers)



Finding rho-independent transcription terminators

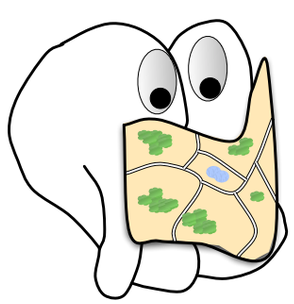


Predicting protein function through network alignment



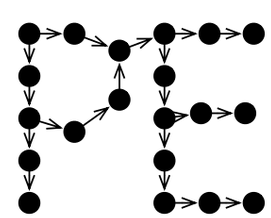
Network phylogenetics

Modeling network evolution

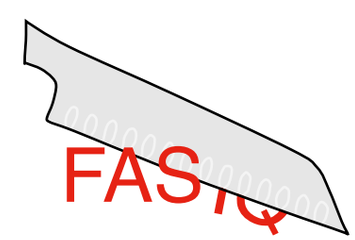


Constructing ribosome footprint profiles

Finding influenza reassortments



Reference-based sequence compression



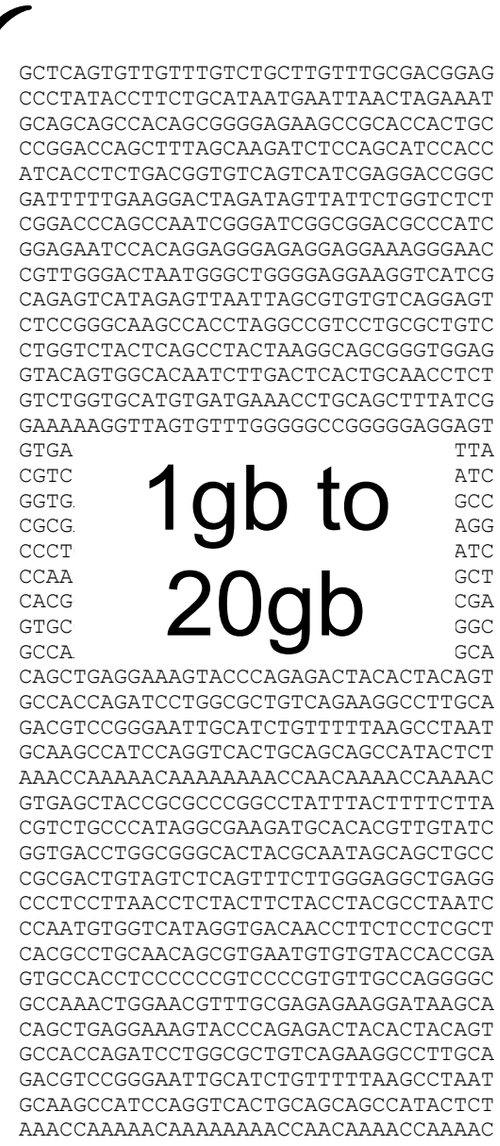
De novo sequence compression

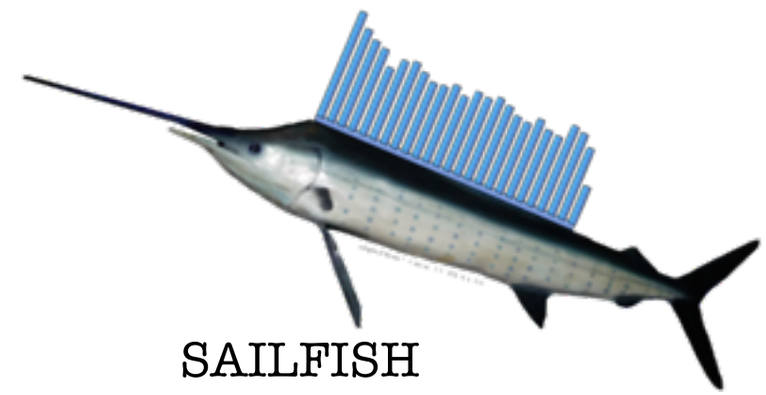
Sailfish: Ultra-fast Gene Expression Estimation

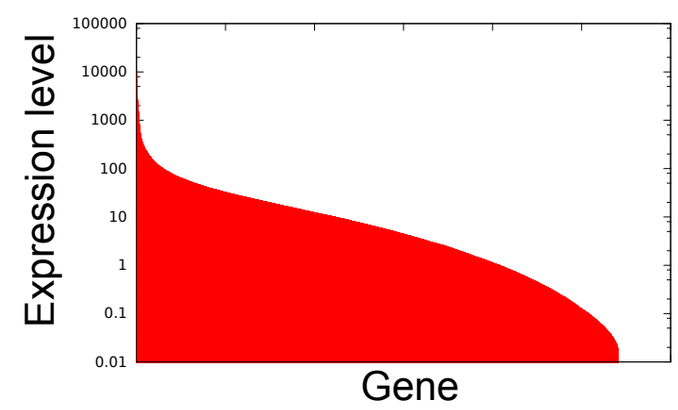
- Measuring gene expression is a fundamental way to uncover organism response to stimuli & to determine gene function

RNA-seq:

10m to 100m reads sampled from genes expressed during a condition



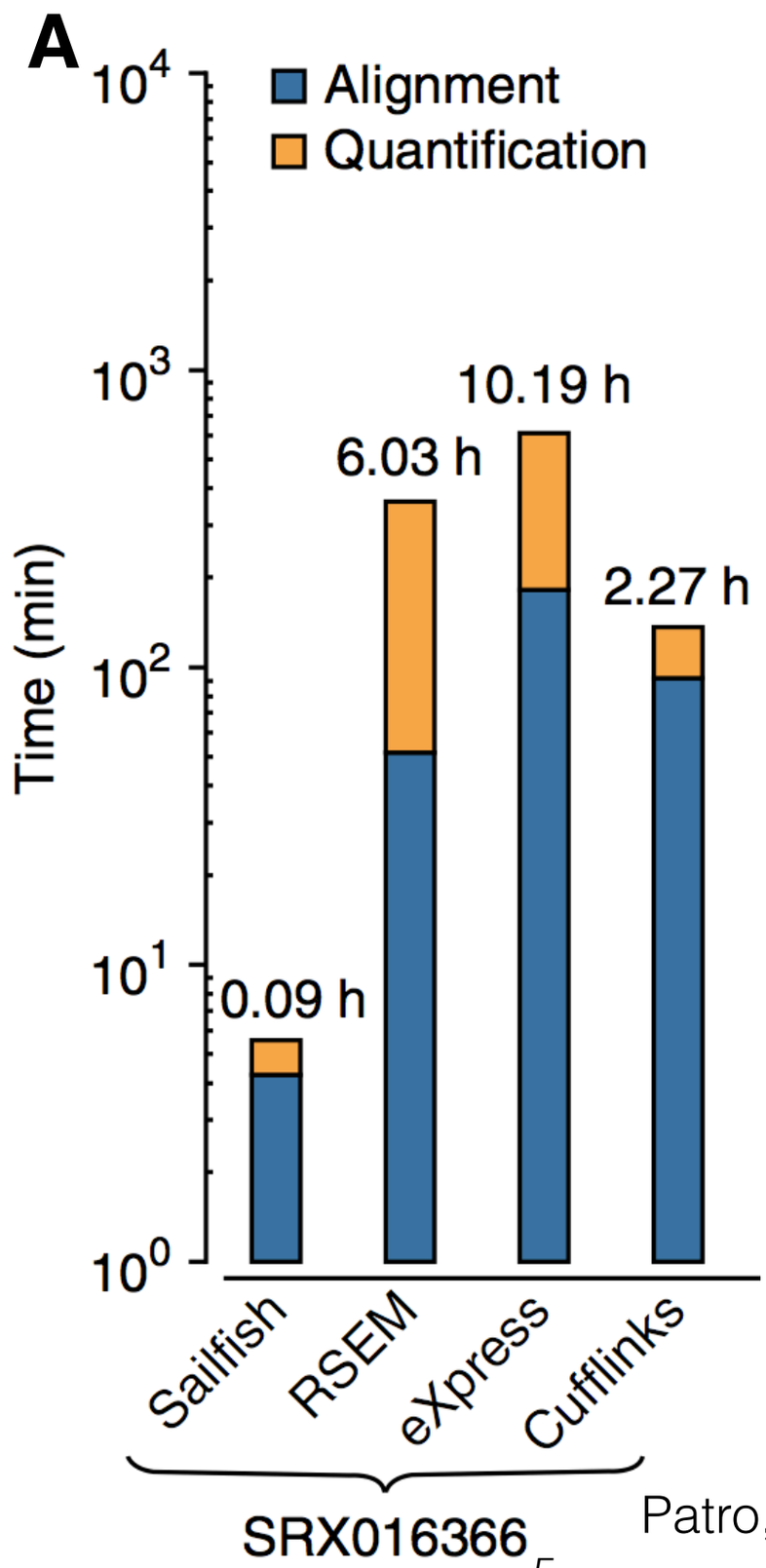


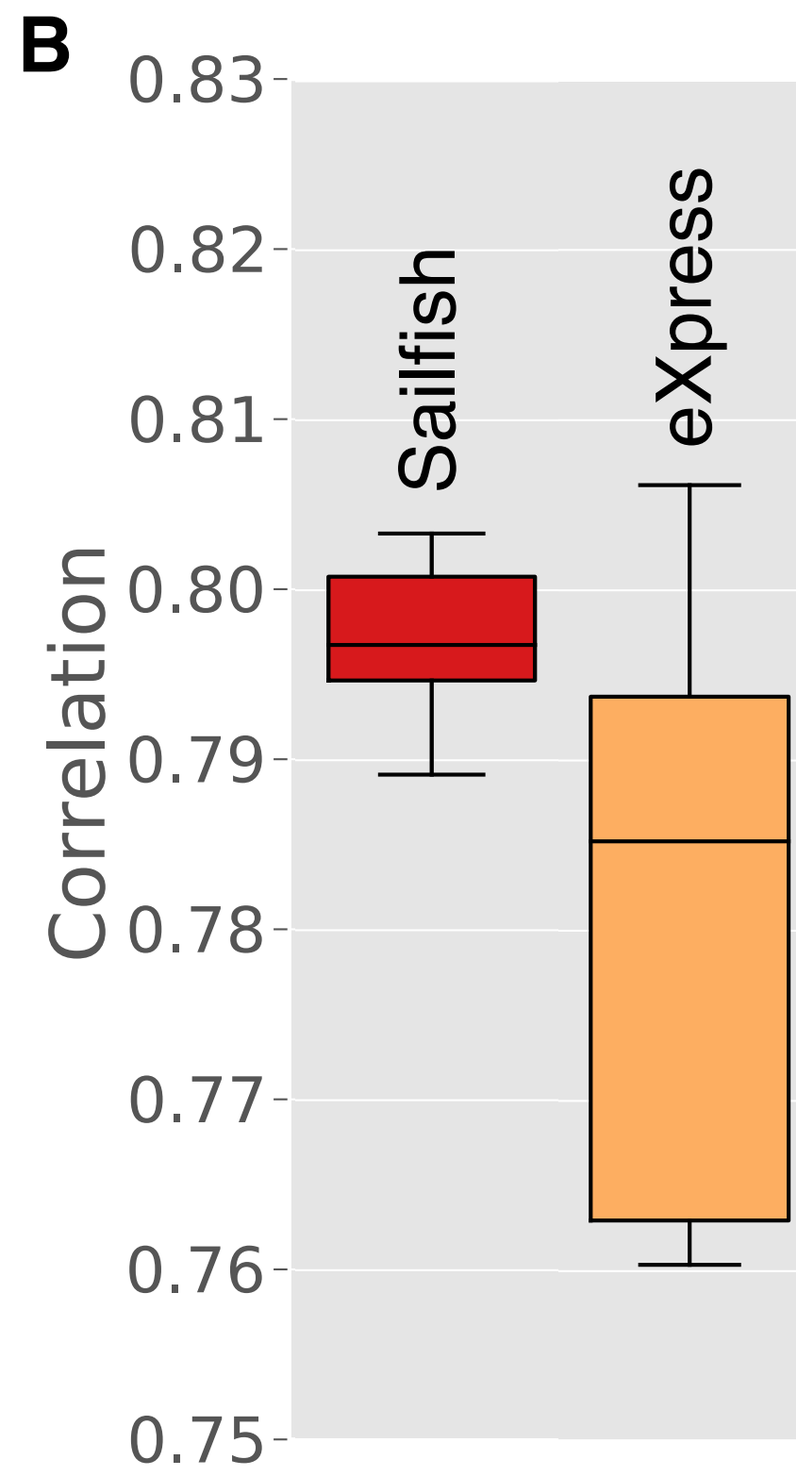


Sailfish quickly determines the relative expression level of genes and their isoforms

Sailfish: Ultrafast Gene Expression Quantification

- Fast expectation maximization algorithm
- Extremely parallelized
- Uses small data atoms rather than long sequences
- More tolerant of genetic variation between individuals



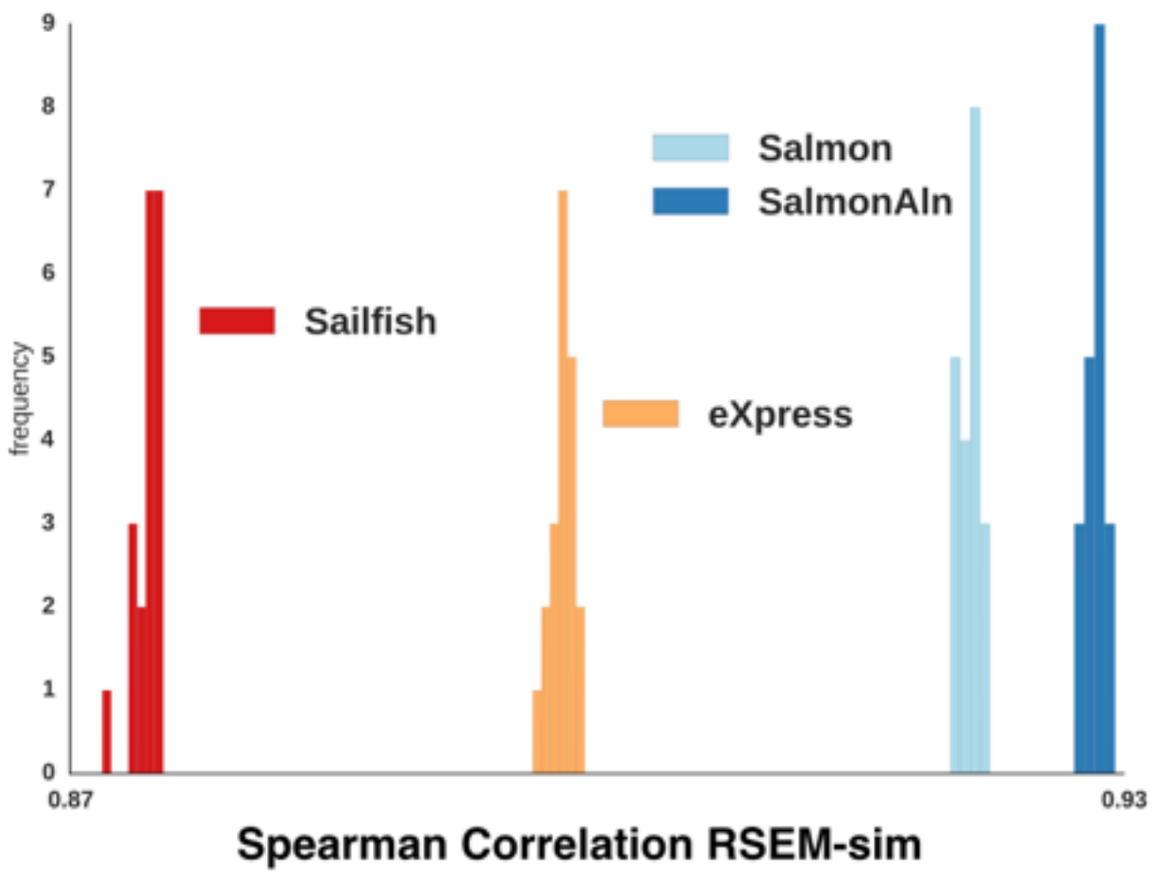


Patro, Mount, Kingsford, *Nature Biotech*, 2014

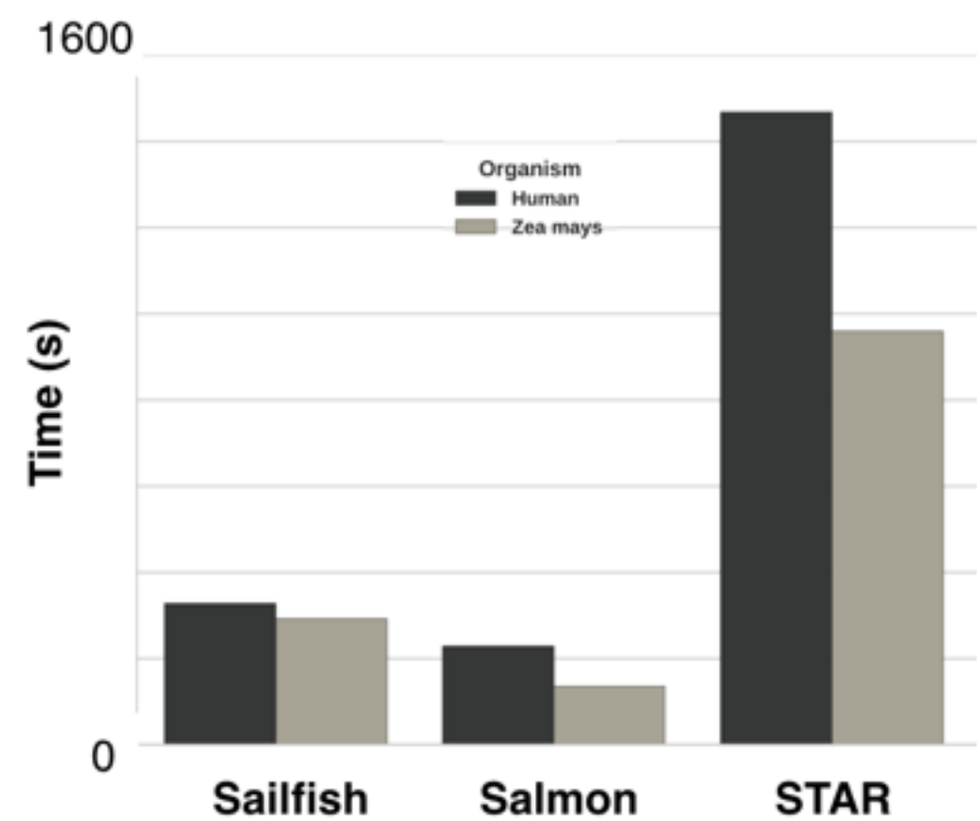
Salmon

- Estimates transcript expression from RNA-seq short reads
- Two-stage streaming variational Bayes / EM
- Novel lightweight alignment algorithms matches reads to transcripts

More accurate



Faster

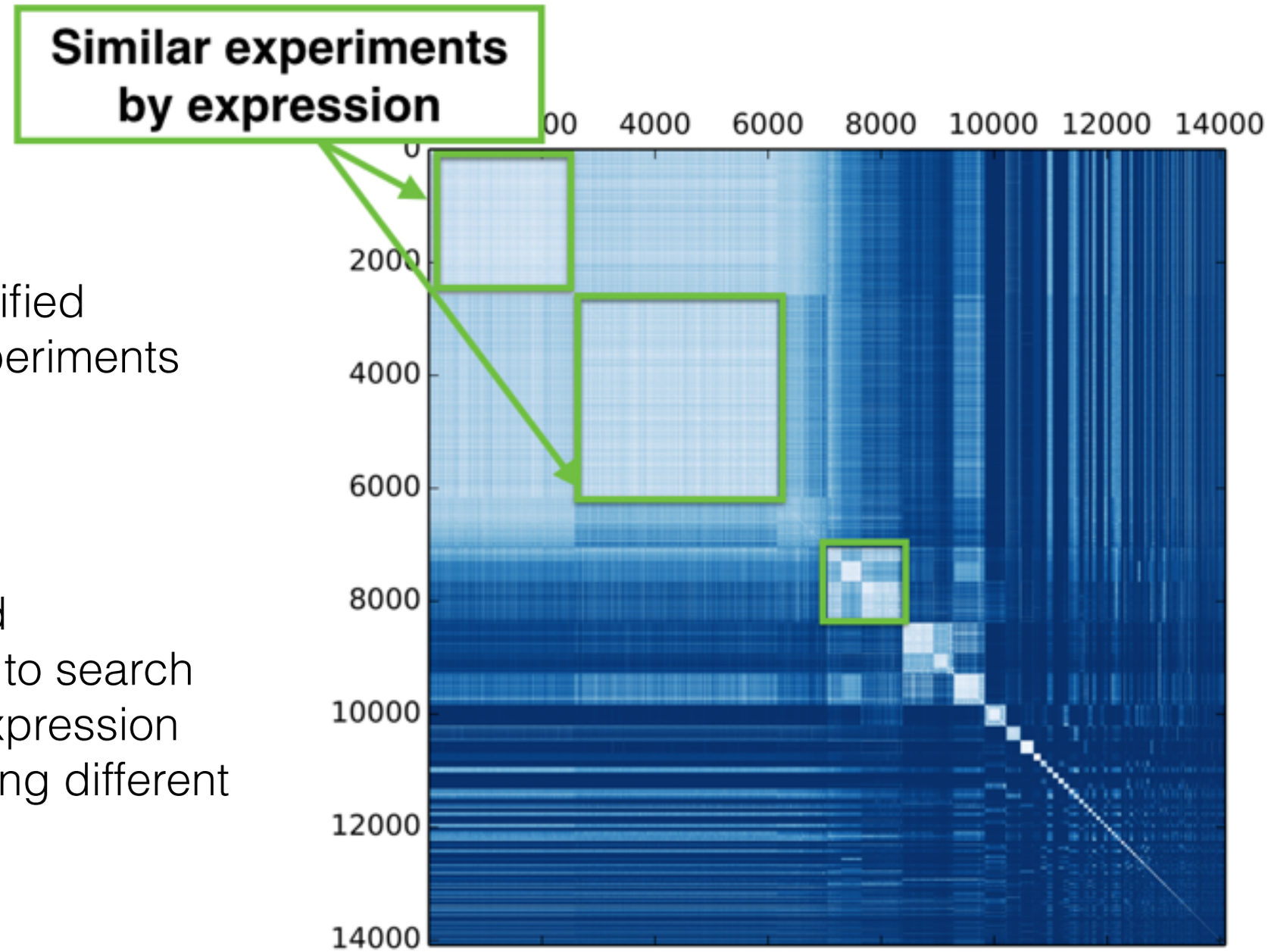


← Better →

Patro, Duggal, Kingsford, http://biorxiv.org/content/early/2015/06/27/021592

“Large-scale Salmon”

- Goal: quantify expression for 100,000 conditions in a consistent way



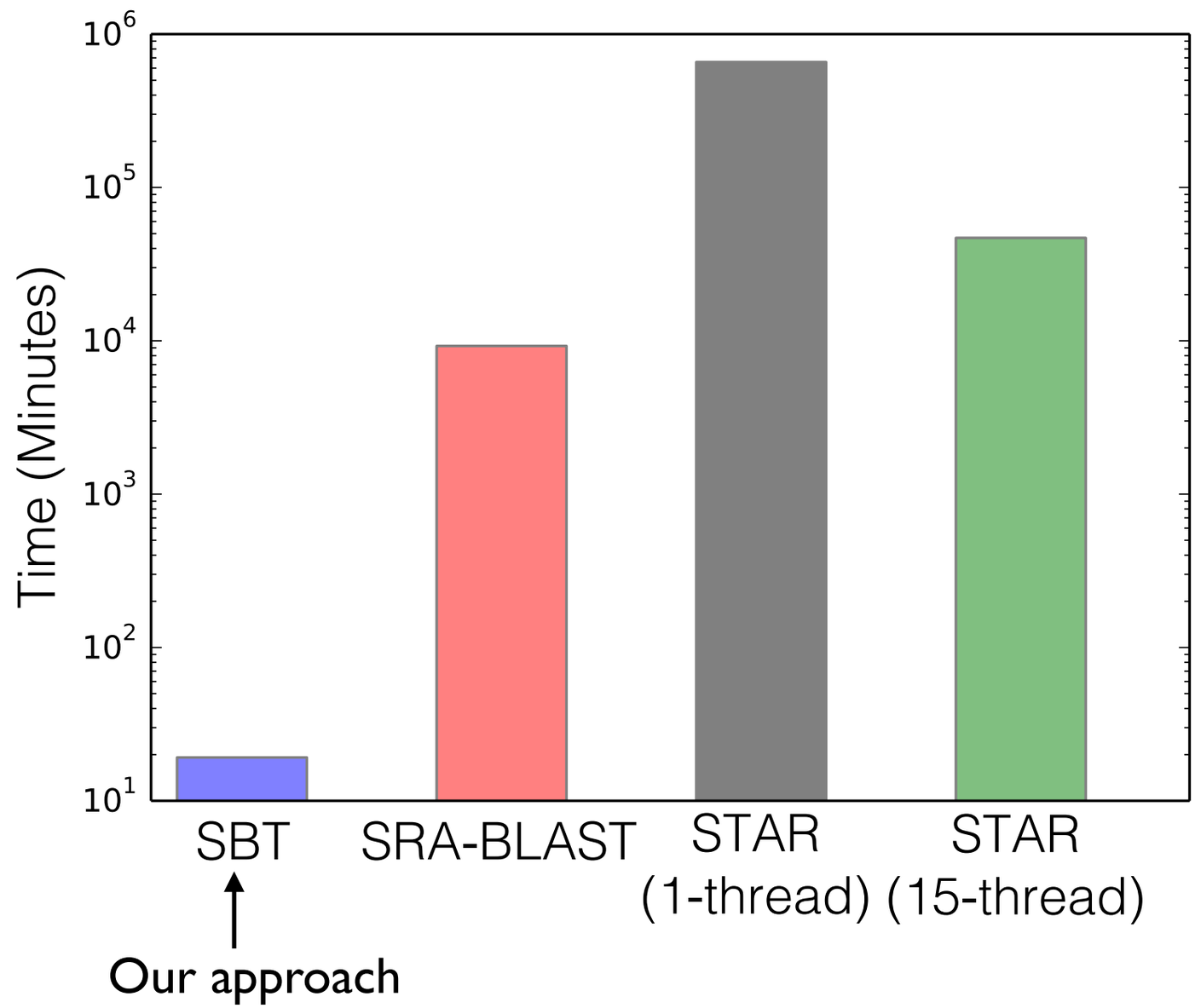
- We’ve quantified 14,000+ experiments currently

- & developed approaches to search for similar expression vectors among different conditions

Finding RNA-seq experiments expressing a given gene

Motivation: Which conditions express a novel gene → hypothesis about the function of that gene.

Time to search 2652 human blood, breast, and brain RNA-seq experiments for a 1000nt gene:

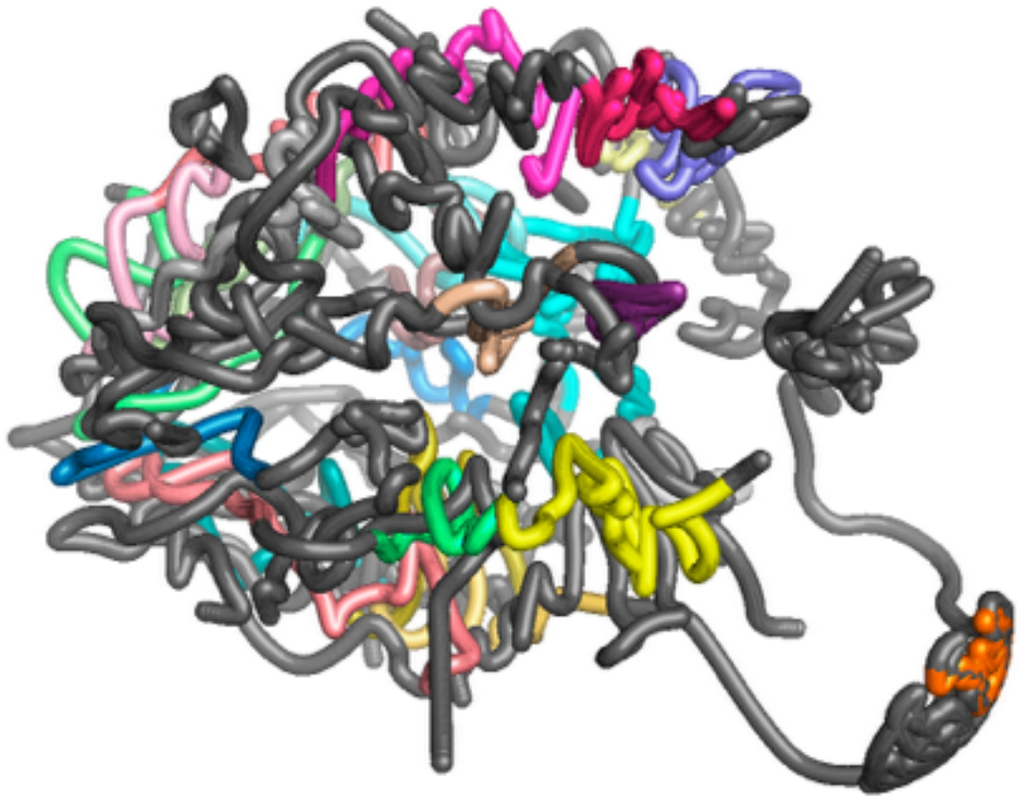


Approach does **not** require that the sequence be a known gene (can search for ncRNA, novel isoforms, new genes).

Things I'm not going to talk about (but ask me!)

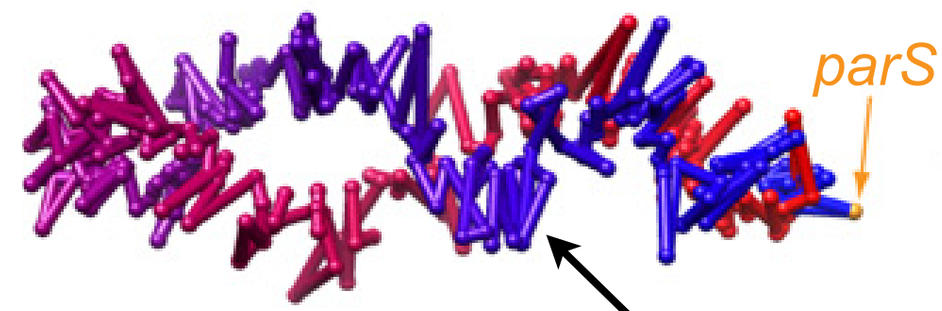
- GHOST - fast, accurate way to compare two large biological networks
- PARANA - parsimonious estimation of network evolution (and prediction of interactions)

Genome Spatial Arrangement

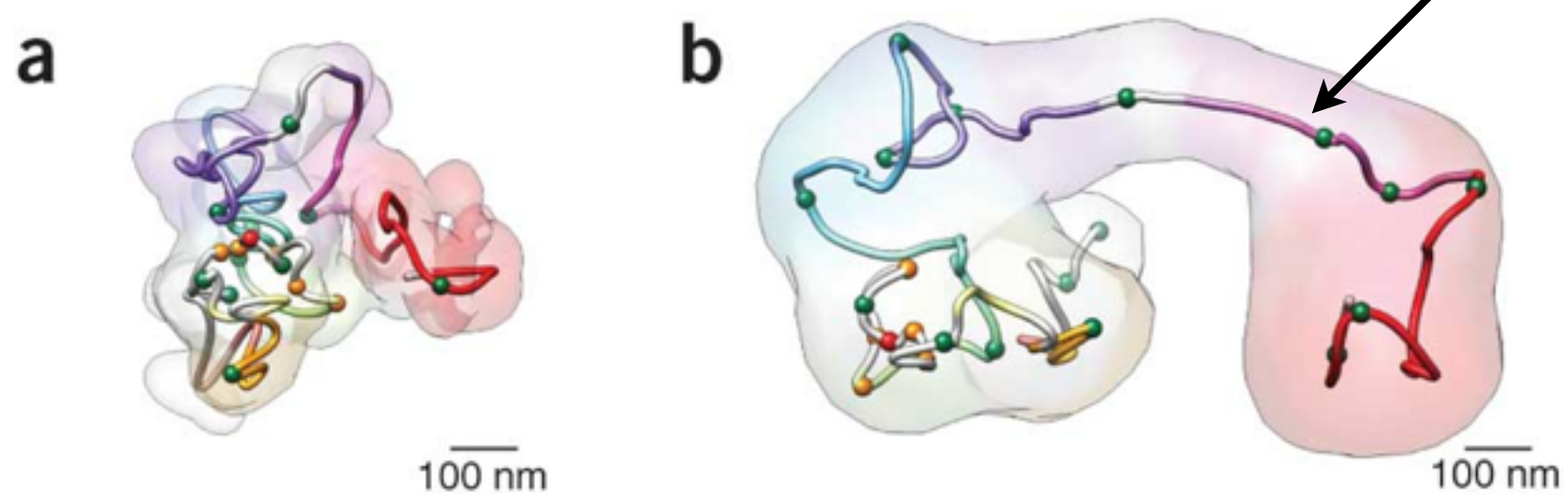


S. Cerevisiae (Duan et al. '10)

Caulobacter crescentus (Umbarger et al. '11)



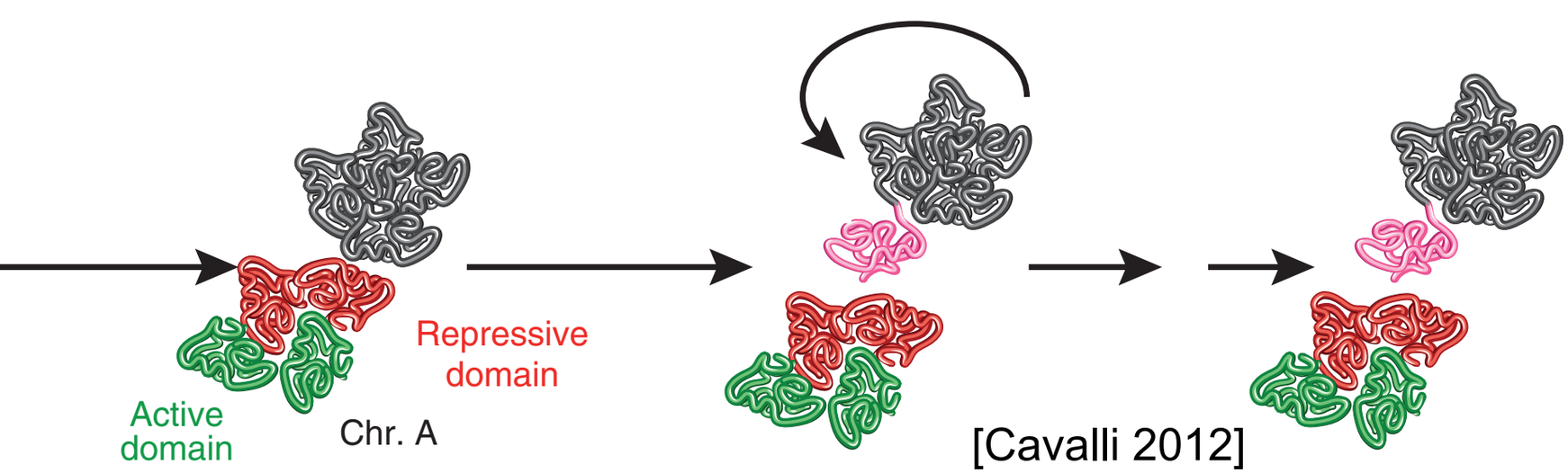
DNA



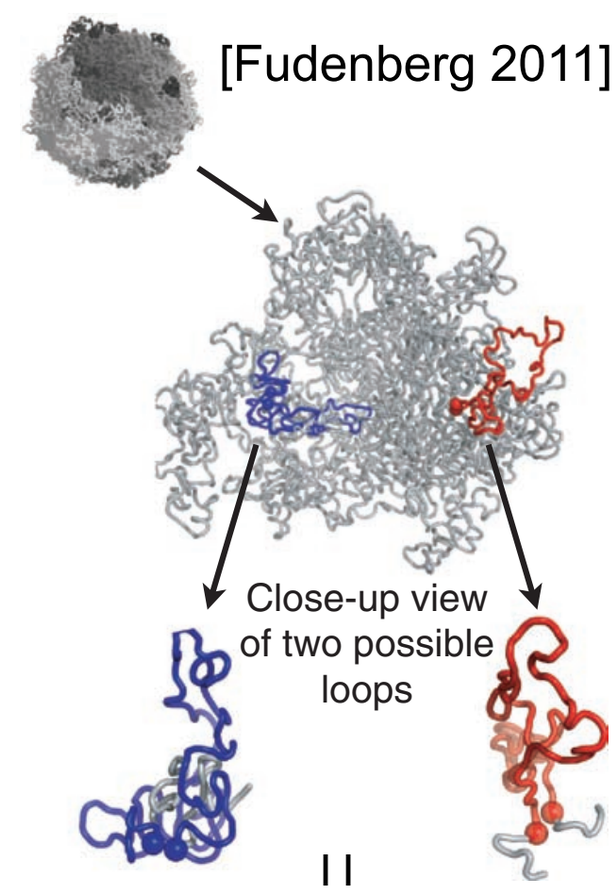
Human healthy vs. cancer (Baù et al. '11)

Chromatin structure is important

- Measured in *Drosophila*, mouse, human,…
- **Implicated in gene regulation and transcription**
- Undergoes important changes during cell development
- Associated with cancer SCNA (e.g. Fudenberg, 2011)

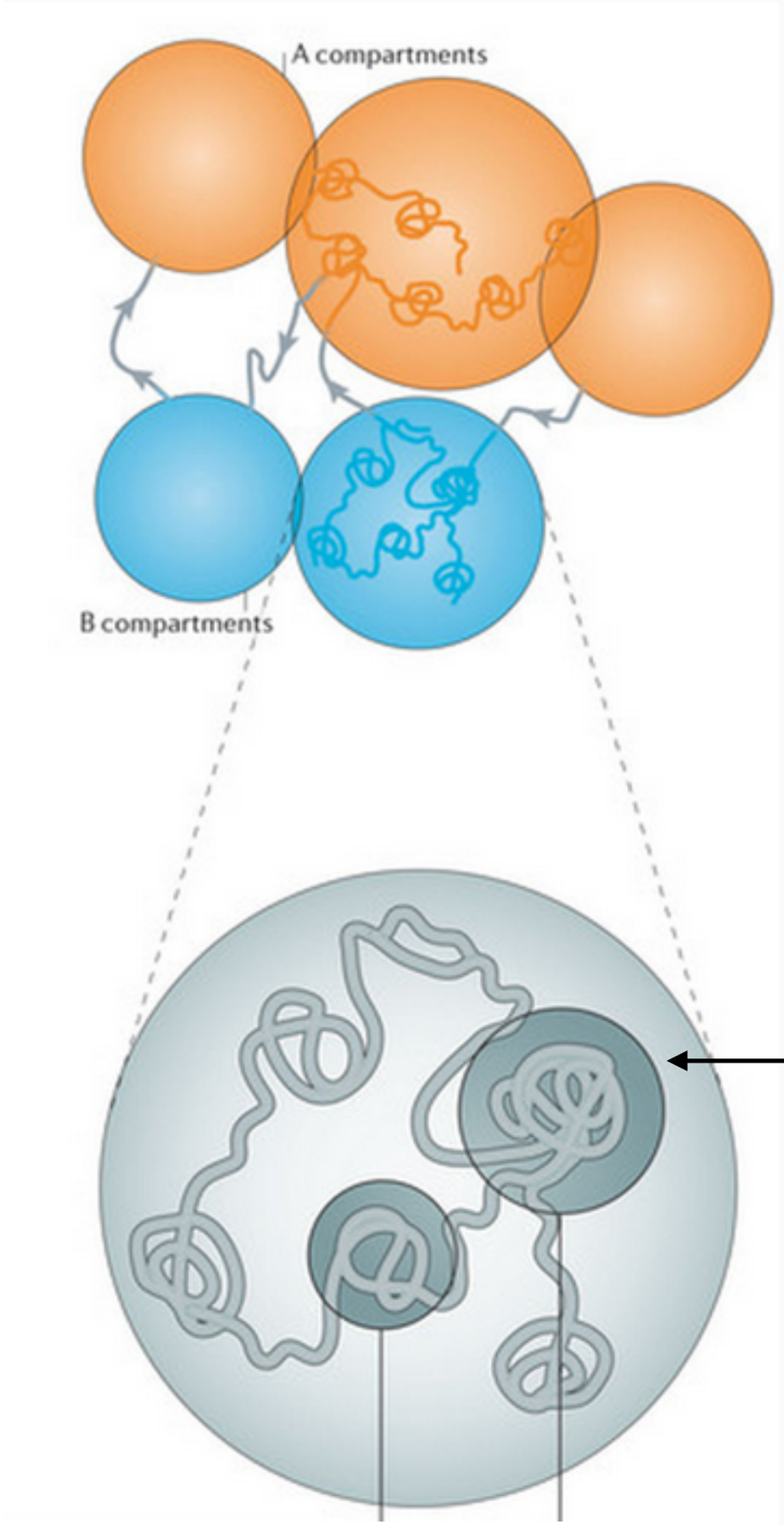


[Cavalli 2012]



[Fudenberg 2011]

“B” compartments
= more dense
regions



“A” compartments
= more open and
loosely compacted

Compact, contiguous
regions = topological
domains (TADs)

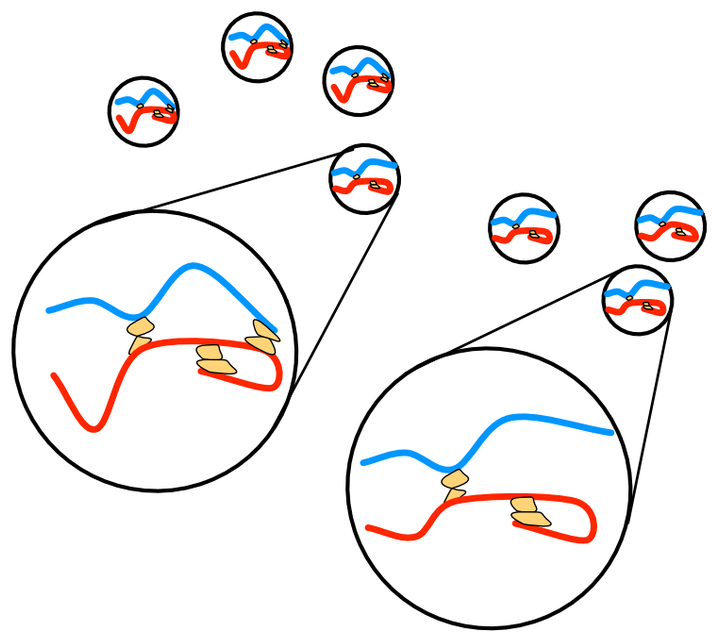
(Dekker et al. ‘13)

Why are TAD's Interesting?

- Stand out as highly-reproducible feature of Hi-C matrices
- Often conserved across species
- Seem to be a key building block of hierarchical organization of chromatin structure
- Play a crucial role in facilitating gene co-regulation and robustness of gene expression

Hi-C: High Resolution, Genome-Wide Structure

Chemically bond spatially close regions of genome across millions of cell nuclei

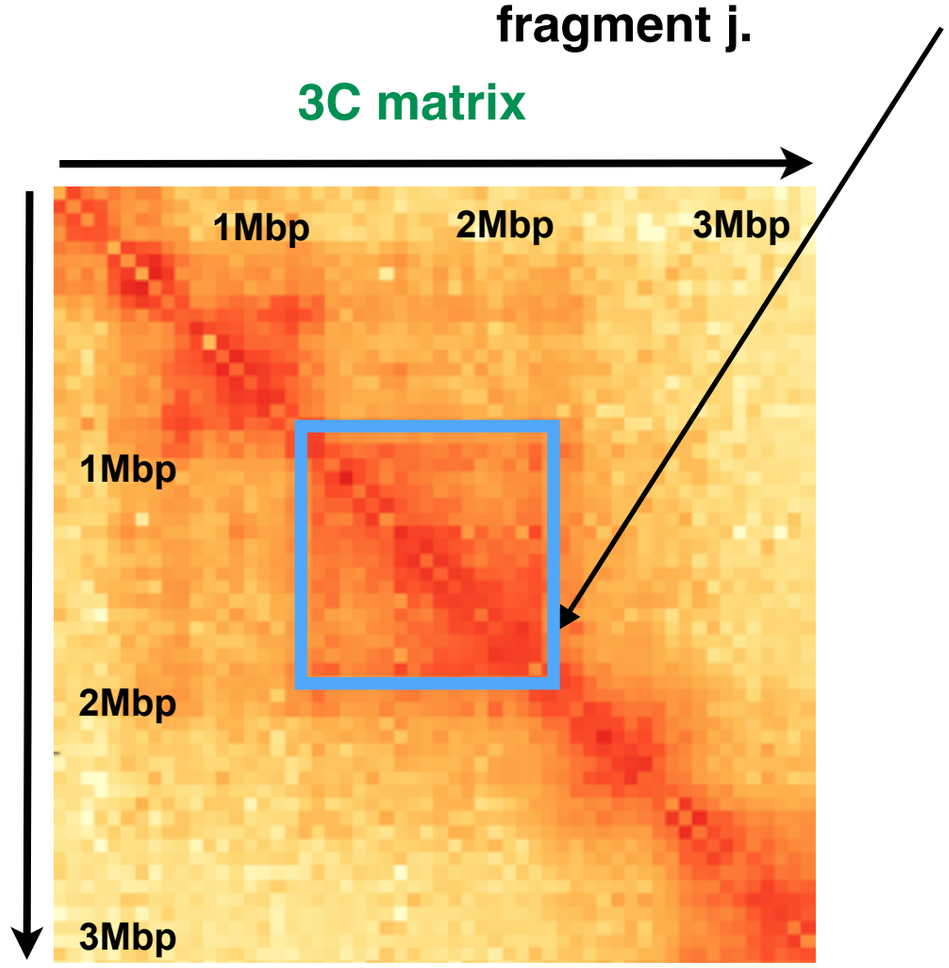


Perform high throughput sequencing to obtain code of nearby regions

Error correct, Normalize, & Filter

(i,j) - # of times DNA at fragment i spatially co-located with DNA at fragment j.

3C matrix



distance is related to 1/ frequency

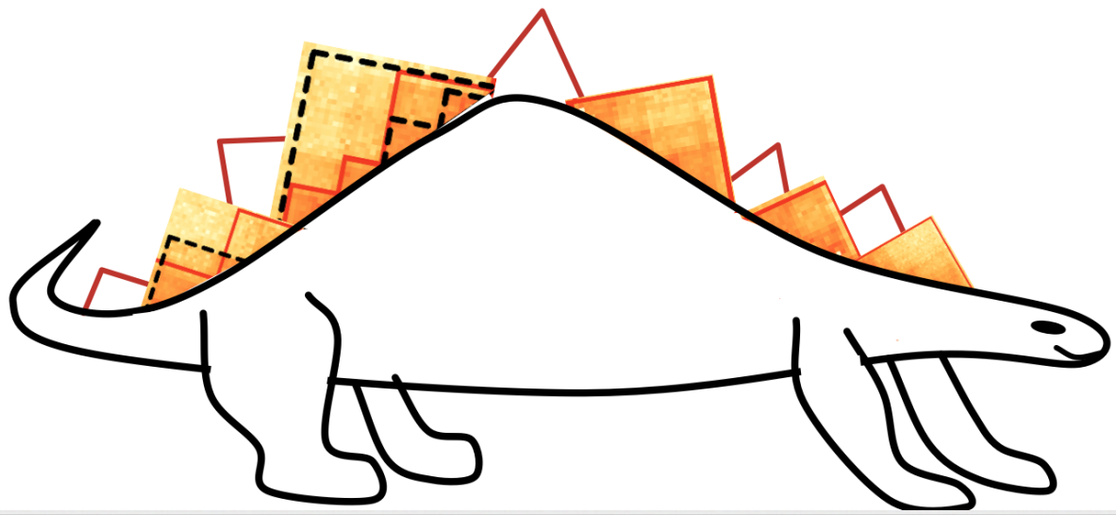
Introduced 2002, whole-genome 2009

Domain-finding Methods

- **Directionality Index HMM (Dixon et al. 2012)**: imbalance between upstream and downstream interactions.
- **Distance-Scaling (Sexton et al. 2012)**: insulation score between upstream and downstream fragments
- **Armatus (Filippova, 2013)**: multiscale domains identified using a interaction density score for the block diagonal.
- **HiCSeg (Levy-Leduc 2014)**: Maximum likelihood formulation to segment Hi-C matrix.
- **Arrowhead (Rao et al. 2014)**: directionality bias at a particular distance *d*. Results in modified contact matrix that looks like it has arrowheads. Heuristically finds domains thereafter.

Armatus

(Filippova, Patro, Duggal, Kingsford. ‘14)



GitHub, Inc. [US] https://github.com/kingsfordgroup/armatus/

This repository Search Pull requests Issues Gist

kingsfordgroup / armatus Unwatch 5

85 commits 2 branches 2 releases 2 contributors

Branch: master armatus / +

Minor modification to Release Creation file

geetduggal authored on May 20 latest commit 50aada0a53

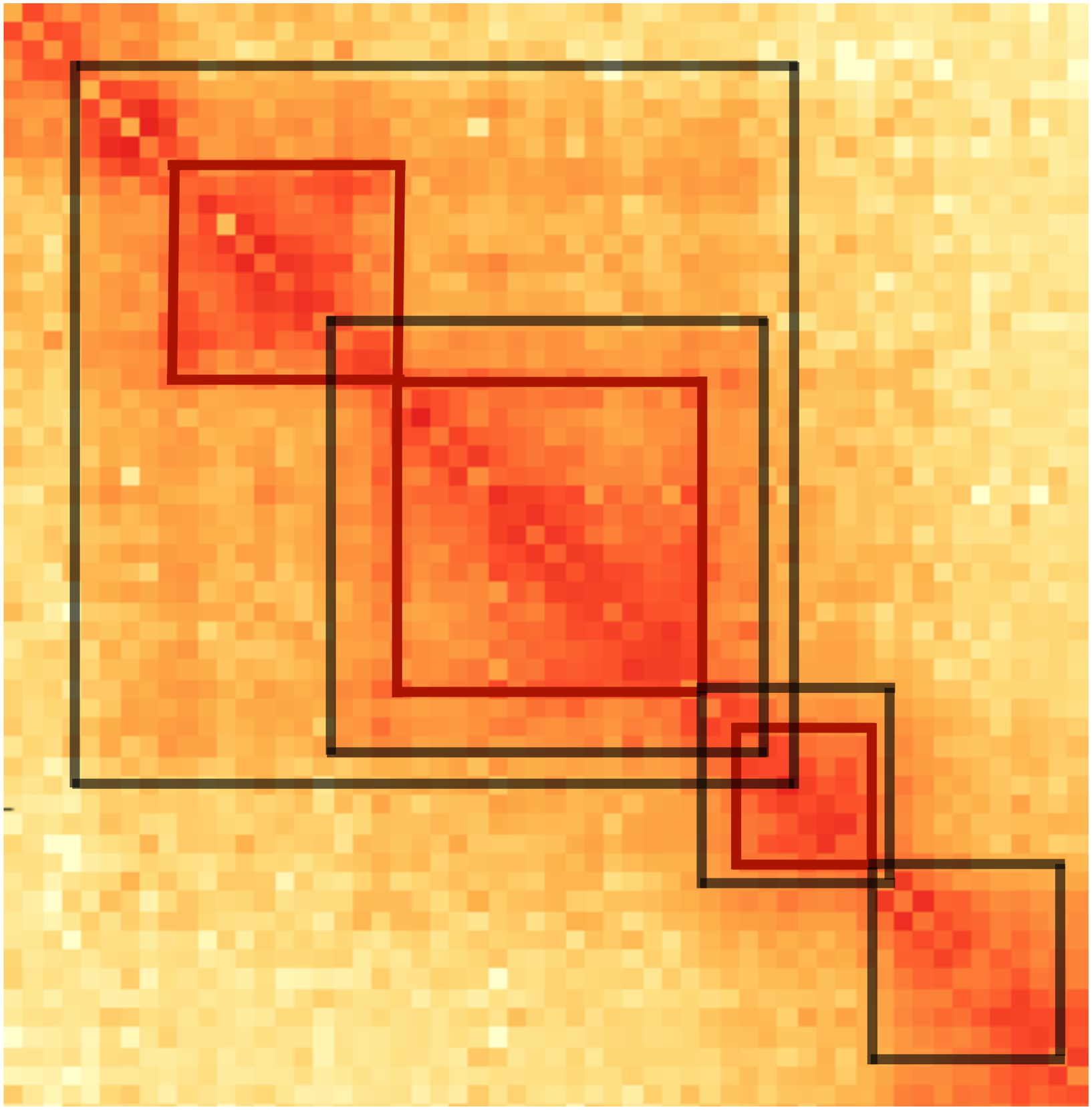
Armatus Features

- First program for **multiscale** analysis of domain structure
- Directly encodes/specifies quality of domain
- Handles uncertainty by generating **multiple near-optimal** solutions
- Order of magnitude **more efficient** than original single-scale analysis
- Efficient enough for highest-resolution data to date
- Requires only a **single parameter**

Domains at Multiple Scales

Dixon et al. domains

alternative domains

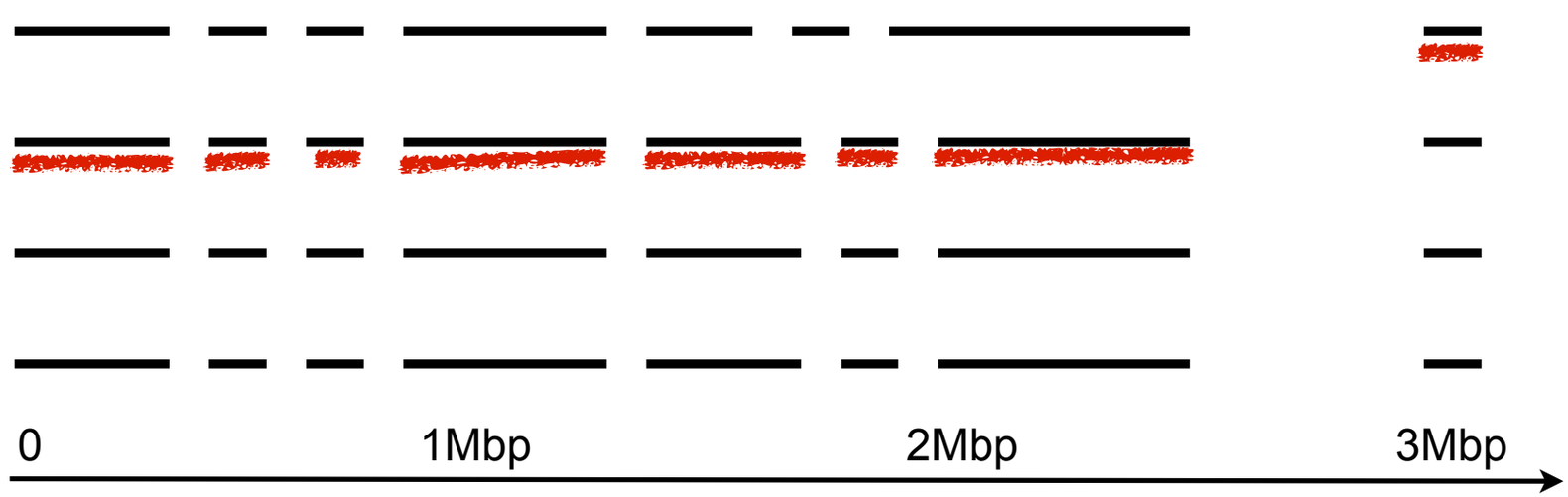


IMR90, chr1

How to find multiscale domains?

1. Find domains: dense regions of high-frequency interactions at different resolutions

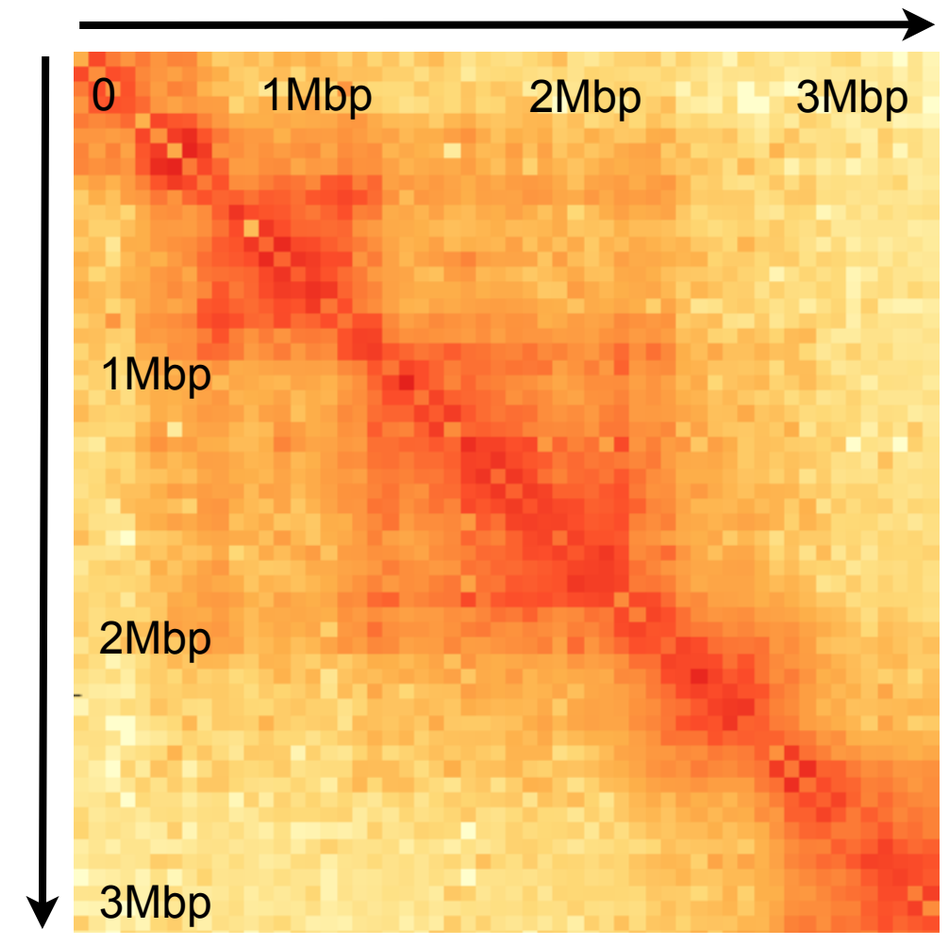
2. Build consensus: pick the most <u>persistent</u> domains to form a single collection



How to find multiscale domains?

1. Find domains: dense non-overlapping square blocks along the diagonal

$$\max \sum_{\text{domains}} q(\text{domain})$$



A - symmetric Hi-C matrix

2. Build consensus: pick domains across resolutions to form a single collection of non-overlapping blocks

$$\max \sum_{\substack{\text{domains at} \\ \text{various scales}}} p(\text{domain})$$

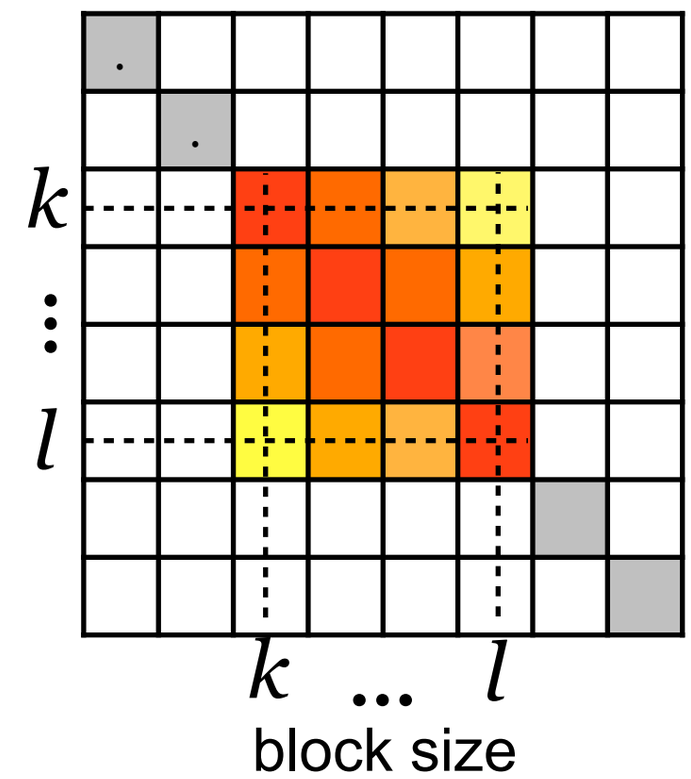
Score dense blocks on the diagonal

block score (can be negative)

$$q(k,l,\gamma) = s(k,l,\gamma) - \mu(\text{size},\gamma)$$

$$\underbrace{s(k,l,\gamma) = \frac{\sum_{g=k}^{l}\sum_{h=g+1}^{l} A_{gh}}{(l-k)^{\gamma}}}_{\text{block weight}}$$

mean weight as a function of block size and resolution



Resolution parameter

block weight $$s(k,l,\gamma) = \frac{\sum_{g=k}^{l} \sum_{h=g+1}^{l} A_{gh}}{(l-k)^{\gamma}}$$

resolution

big domains

small domains

$\gamma = 0:$ denominator becomes 1

$\gamma = 1:$ $|E|/|V|$ as used in [Goldberg 84]

$\gamma = 2:$ $|E|/\binom{|V|}{2}$ similar to weighted edge density

Resolution-Specific DP

End in a non-domain

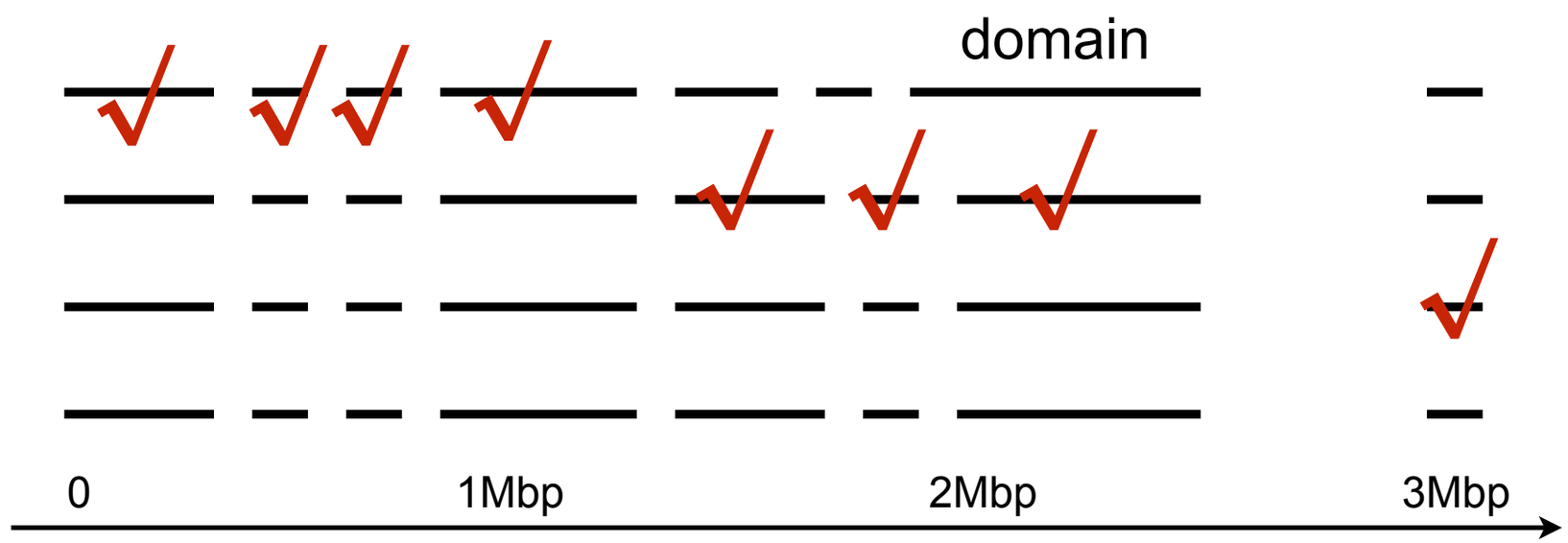
$$\mathsf{OPT}'_1(l) = \max \begin{cases} \max_{k<l}\{\mathsf{OPT}_\mathsf{D}(k-1)\} \\ \mathsf{OPT}_\mathsf{D}(l), \end{cases}$$

End in a domain

$$\mathsf{OPT}_\mathsf{D}(l) = \max_{k<l}\{\mathsf{OPT}'_1(k-1) + q'(k,l,\gamma)\},$$

$$q'(k,l,\gamma) = \begin{cases} q(k,l,\gamma) & \text{if } q(k,l,\gamma) > 0 \\ -\infty & \text{otherwise.} \end{cases}$$

Building a consensus of domains

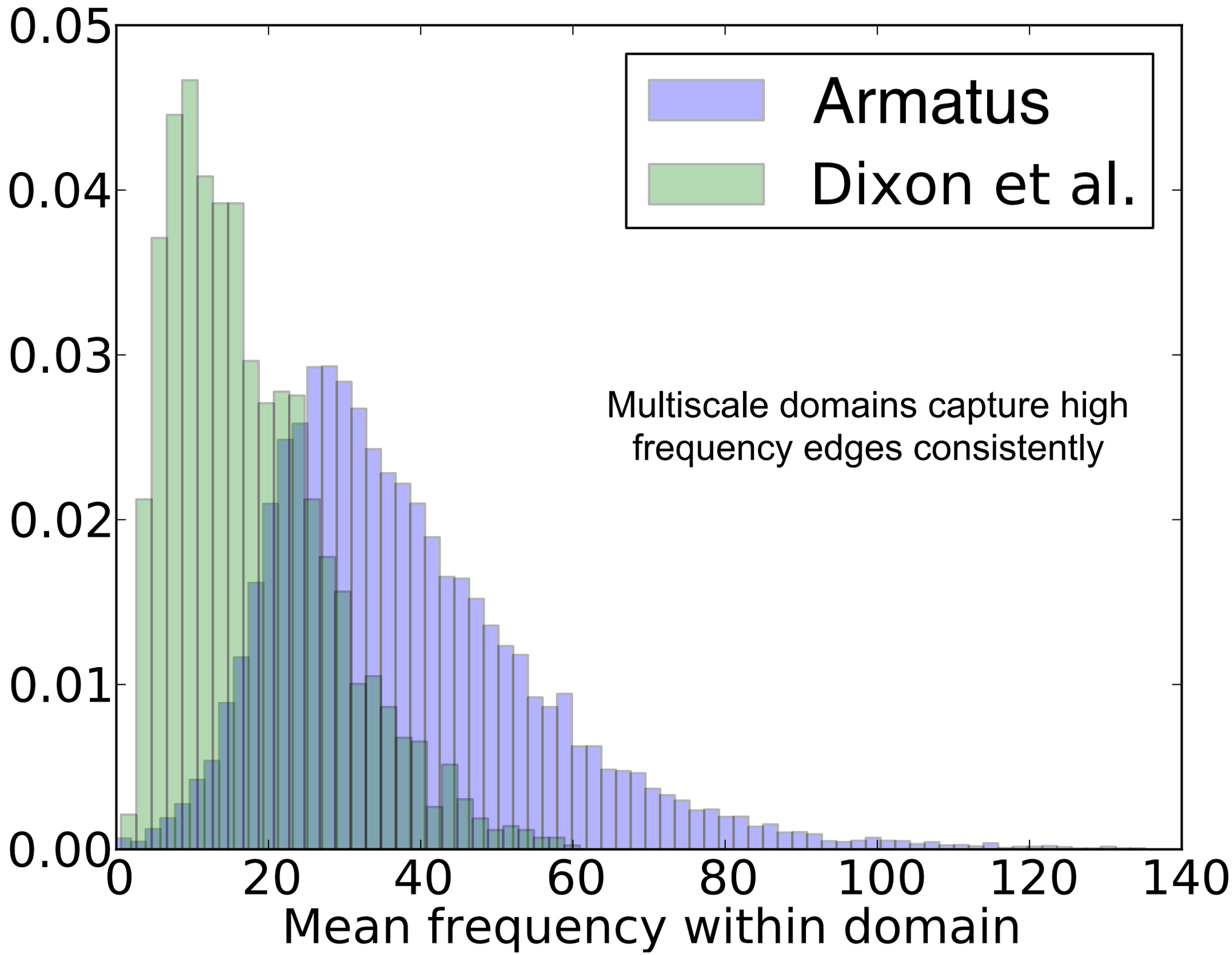


domains = intervals, occurrence = weight

Weighted interval scheduling

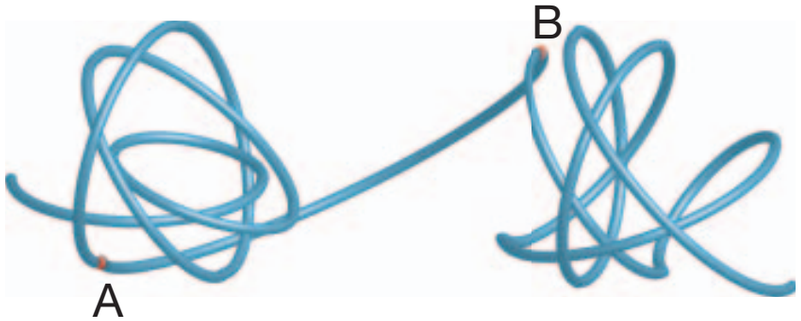
$$\mathsf{OPT_C} = \max \begin{cases} \mathsf{OPT_C}(j-1) & \text{mark } j \text{ as non-domain} \\ \mathsf{OPT_C}(c(j)) + p(a_j, b_j, \Gamma) & \text{extend domain} \end{cases}$$

Distribution of mean interaction frequency

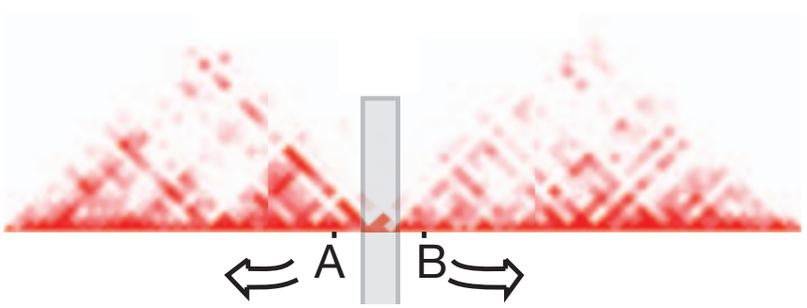


Enrichment for structure-related genomic signals in the boundaries

[Dixon 2012]



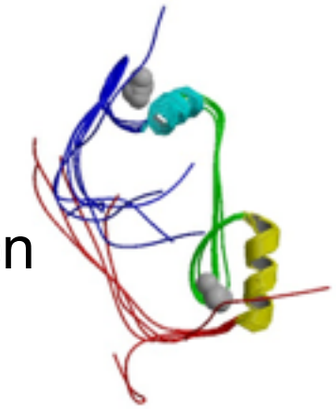
Interactions upstream



boundary - a stretch of DNA between domains, 40-400Kbp

CTCF

- transcriptional regulation
- insulator activity
- regulation of chromatin architecture

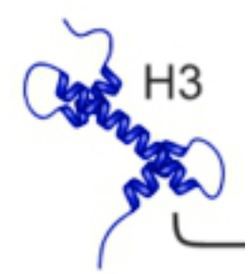


[PDB]

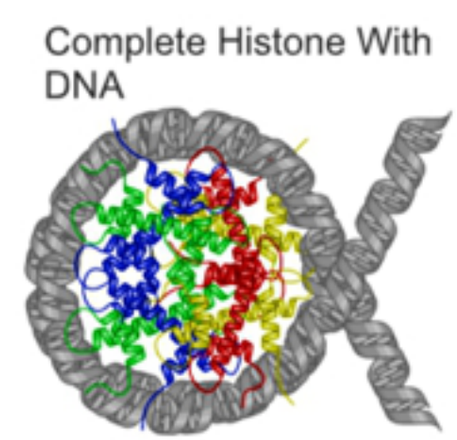
H3K27ac

- chromatin structure in eukaryotes
- form nucleosomes
- H3 most extensively modified

H3K4me3

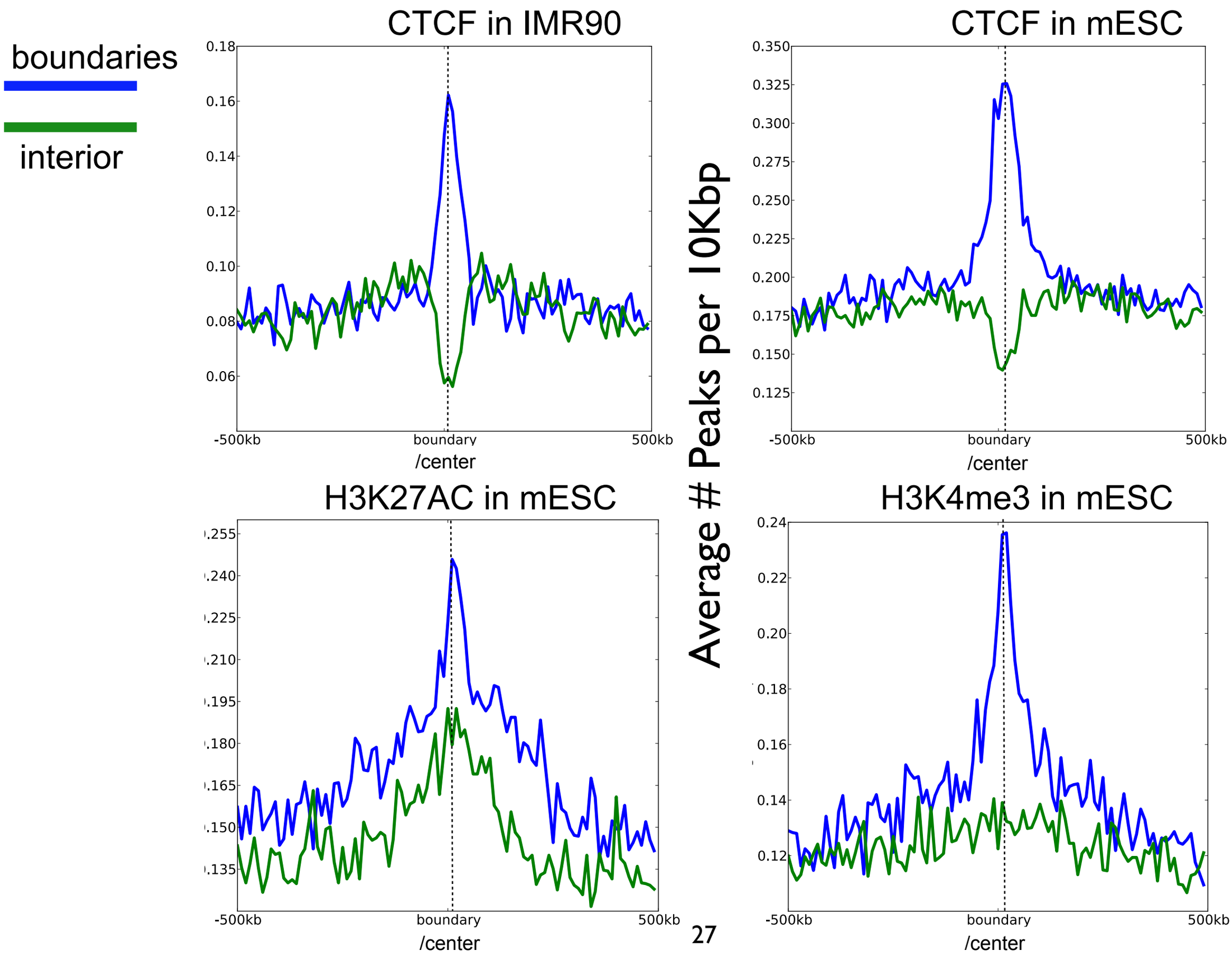


transcription activation/repression



[Wheeler@Wikipedia]

Enrichment for chromatin marks



More functional peaks in multiscale boundaries

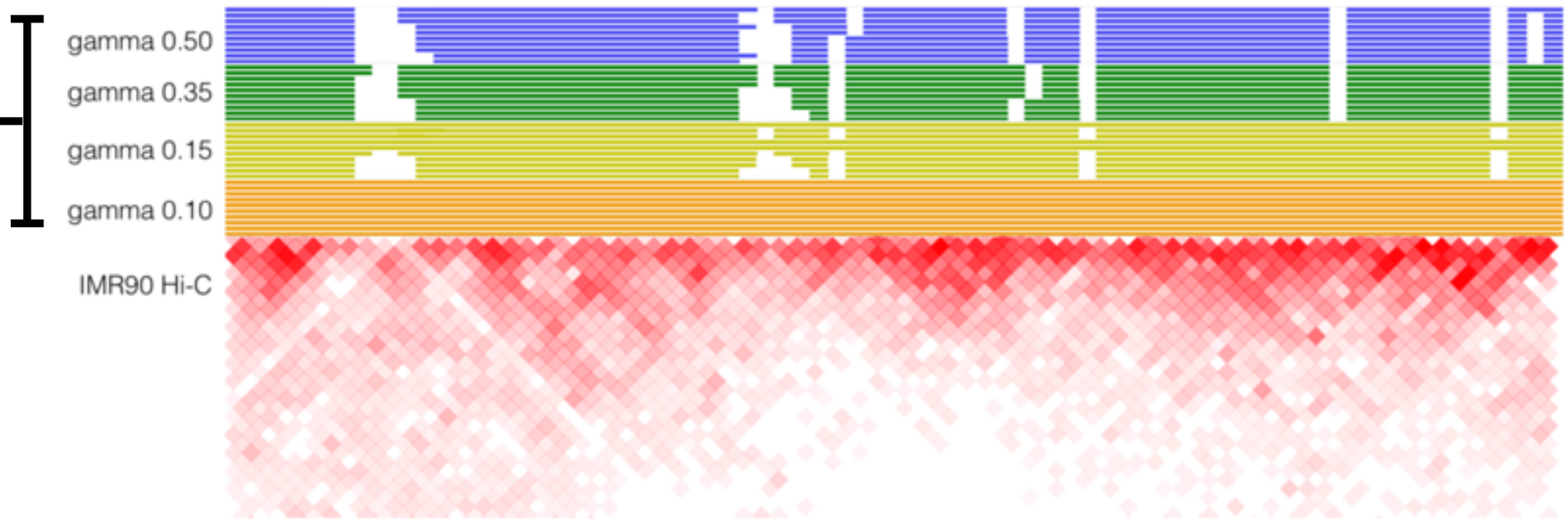
Signal	Boundaries (Dixon)	Boundaries (Armatus)
CTCF (IMR90)	20%	**44%**
CTCF (mESC)	33%	**72%**
H3K4me3 (mESC)	30%	**60%**
H3K27ac (mESC)	23%	**43%**

%boundaries with at least one peak

Also: see peaks less often *within* multiscale domains

Analyses Enabled by High-quality Domains

Multiscale Domains are Hierarchically Organized



Collect all optimal and near optimal-domains across scales into one set

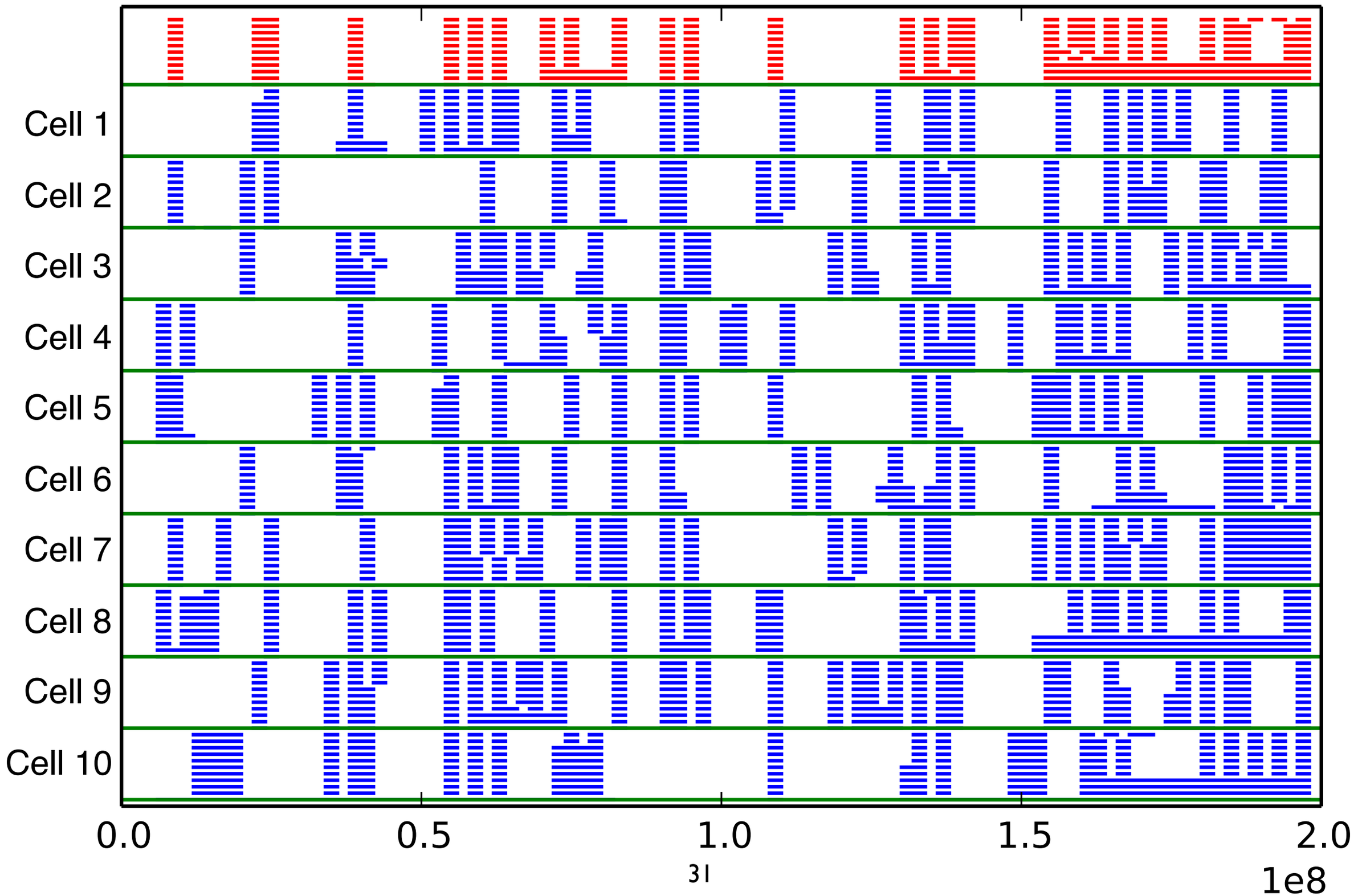
Determine the percentage of all sufficiently different domain pairs di, dj where di is *completely* contained within dj or vice-versa.

95% of all sufficiently different domain pairs are hierarchically organized.

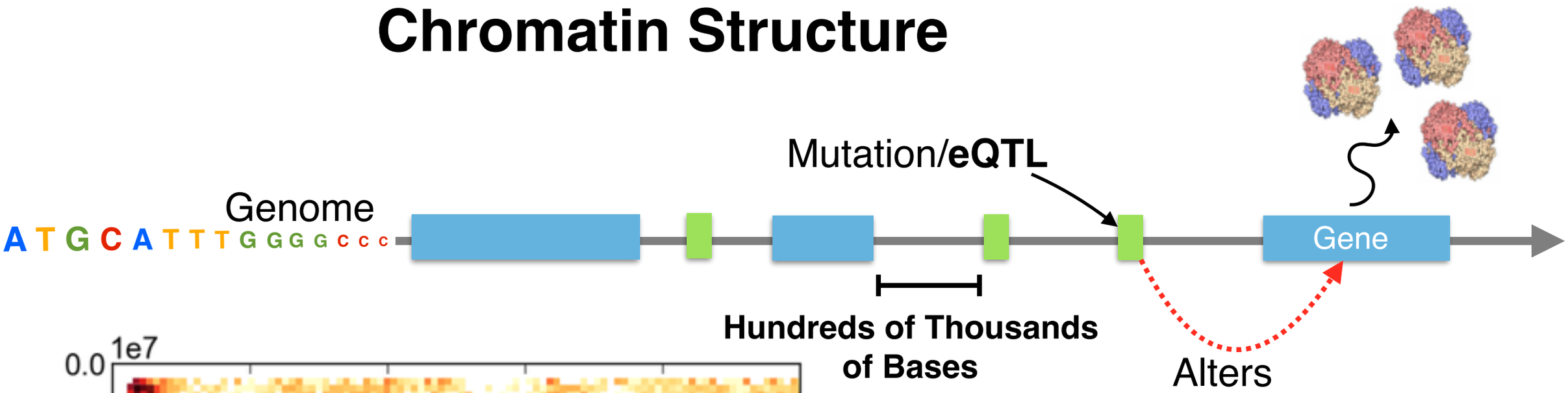
70% of re-shuffled domains are hierarchically organized.

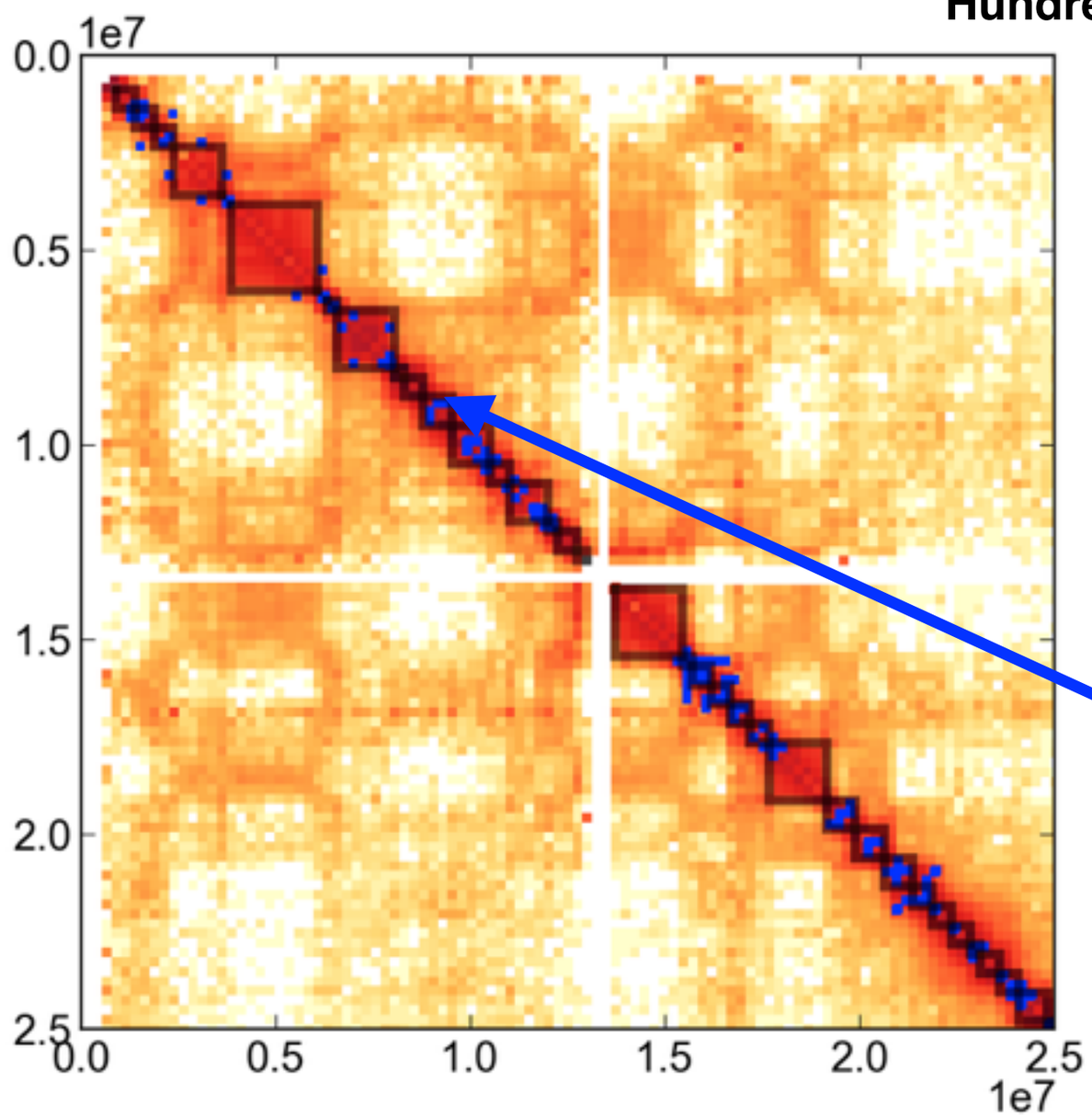
Hierarchy Holds in Single-Cell Data Too

(data from Nagano et al., 2013)



First Genome Wide Analysis Relating eQTLs to Chromatin Structure





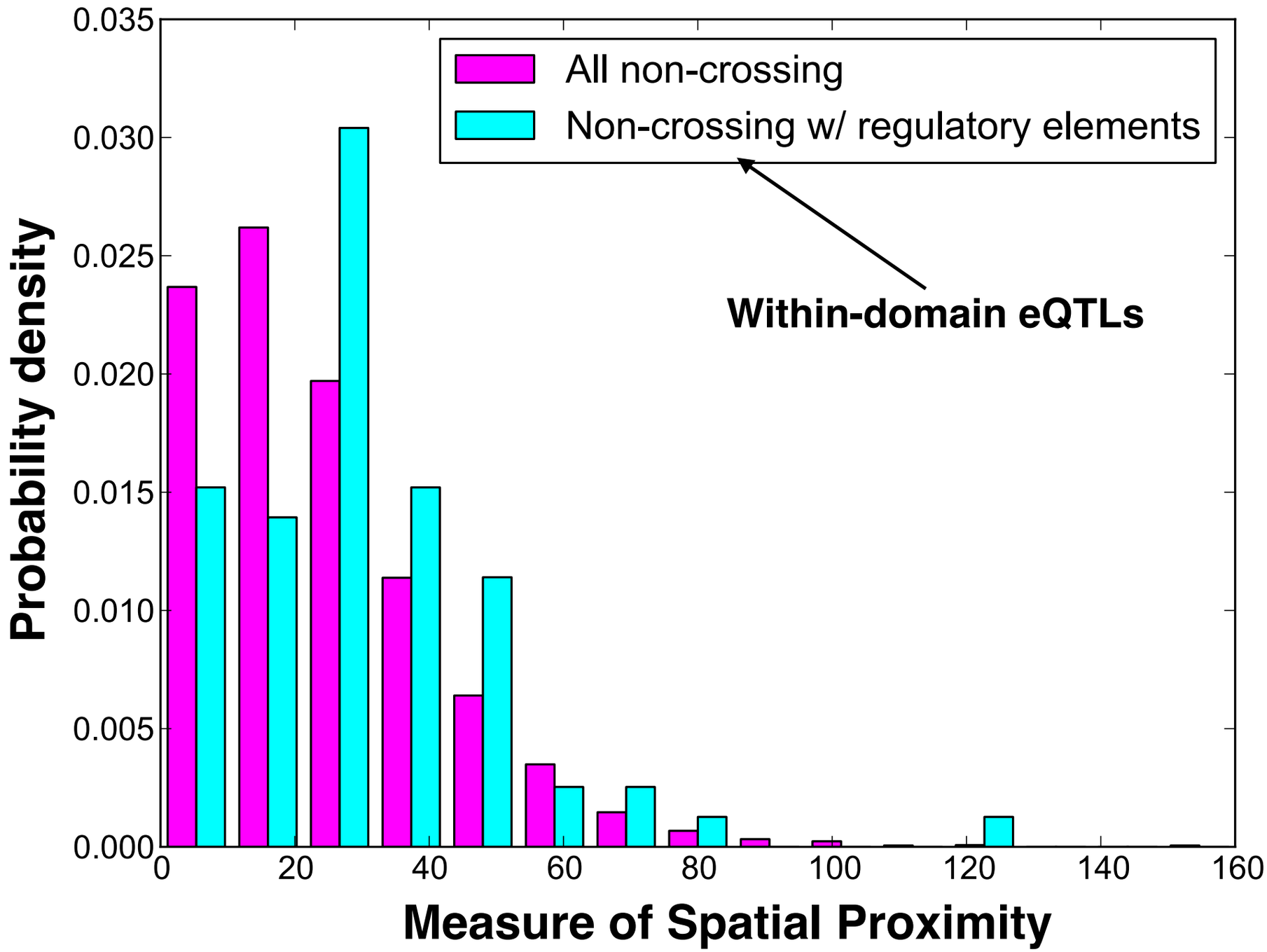
Mutations tend to be spatially close to their target genes.

Occur at the boundaries of domains

Mutation-gene pair

(Duggal, Wang, Kingsford, *NAR*, 2014)

eQTLs Overlapping Regulatory Elements are Surprisingly Spatially Close to their Target Genes



(Duggal, Wang, Kingsford, *NAR*, 2014)

Generative Model for Domain Formation From Histone Marks

- GM log likelihood function

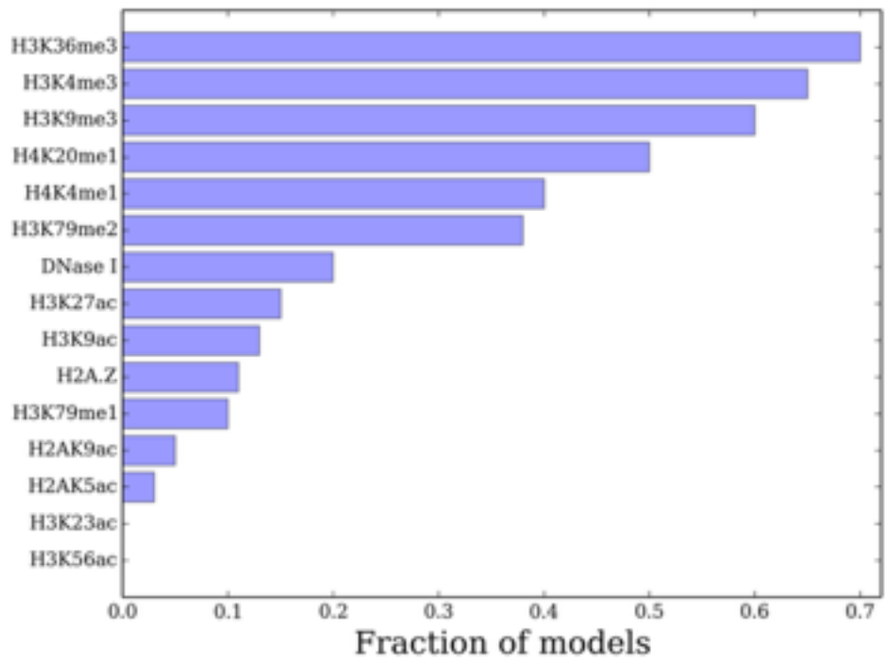
$$\underset{D}{\operatorname{argmax}}\ \log(\,P(D \mid W,H)) \;=\; \sum_{d=[s,e]\in\overline{D}} r_{se}x_{se} + \sum_{v\in V} E_v^e y_v$$

$$\overline{D} = \{[s,e] \mid s,e \in V, e - s \geq 1\}$$

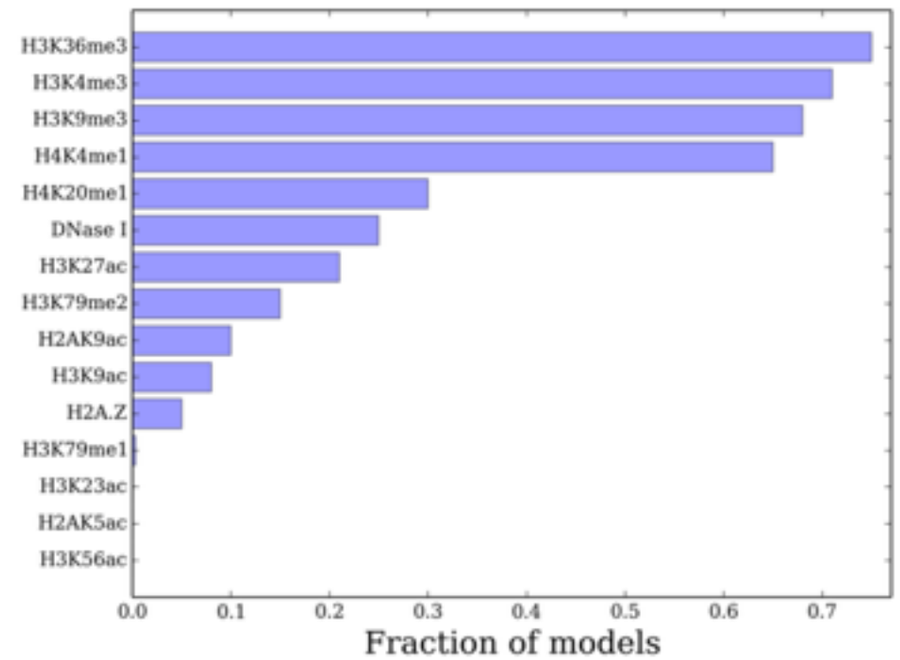
$$r_{se} = E_s^b + E_e^b + \sum_{v=s+1}^{e-1} E_v^i$$

- *x* and *y* are indicator functions for when solution contains [*s*,*e*] and *v* not assigned to domain, respectively

Generative Model of Domain Boundaries From Genomic Markers



(a) human IMR90



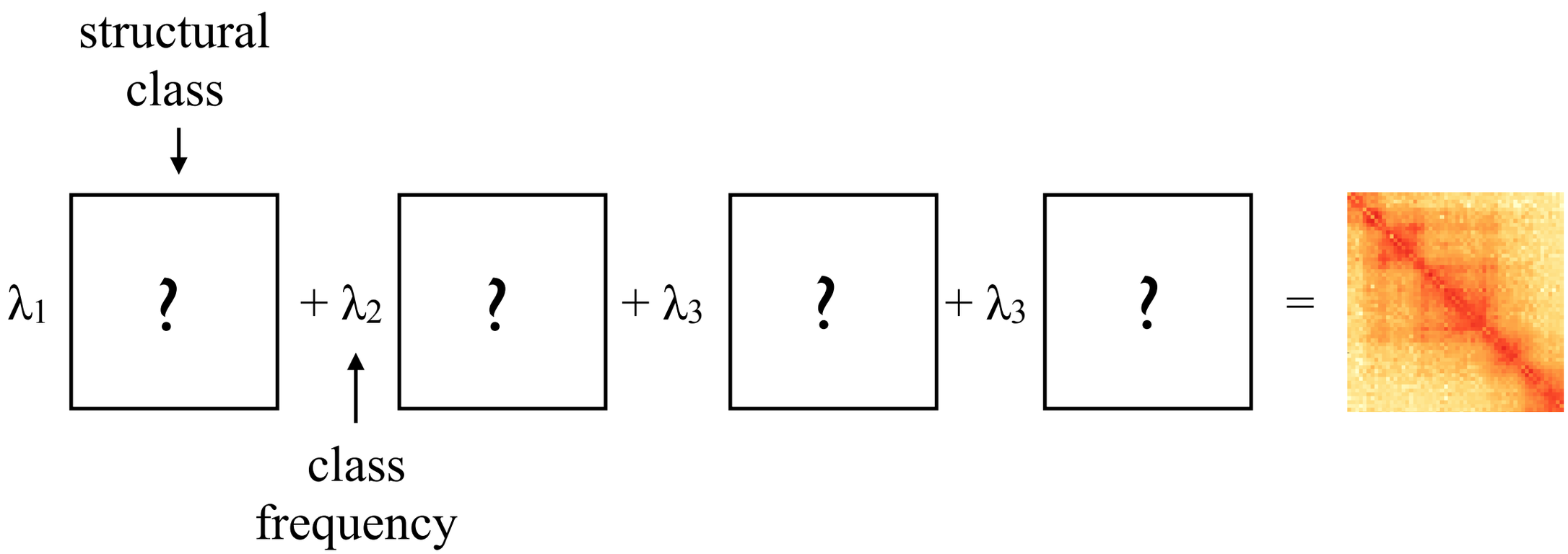
(b) human ES

Table 1: Normalized coherence scores of various marker subsets

Allowed modifications (human IMR90 to IMR90)	Coherence score (Normalized)
28 histone modifications + Concave + Nonnegative *	1.00
28 histone modifications + Concave	0.99
28 histone modifications	0.97
H3K4me3, H3K79me2, H3K27ac, H3K9me3, H3K36me3, H4K20me1	0.94
H3K36me3, H3K4me1, H3K4me3, H3K9me3 + Concave + Nonnegative	0.94
H3K36me3, H3K4me1, H3K4me3, H3K9me3 + Concave	0.93
H3K36me3, H3K4me1, H3K4me3, H3K9me3	0.92

(Sefer, Duggal, Kingsford, WABI 2015)

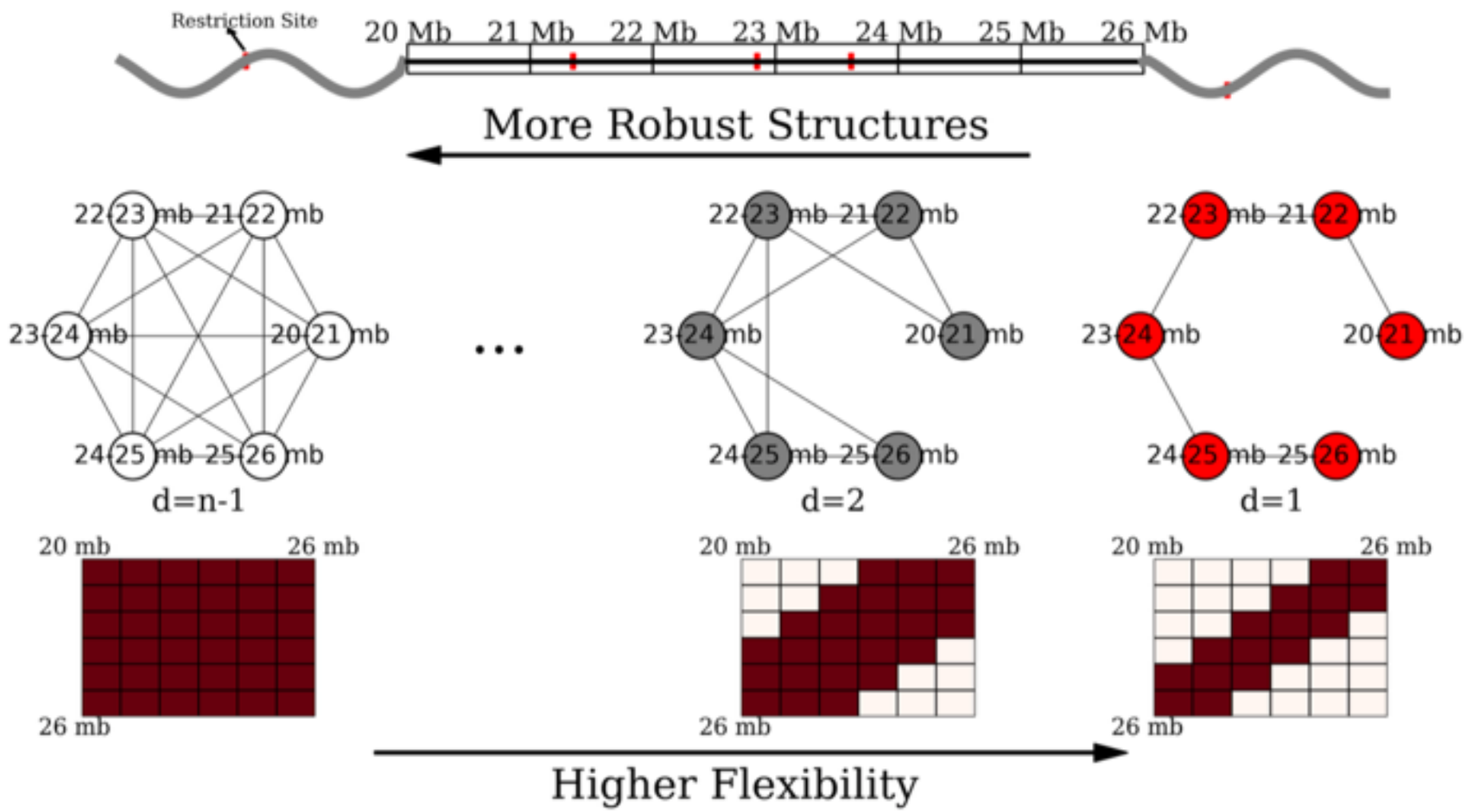
Deconvolution: Estimating Structural Classes From Population Hi-C



- Assume each class composed of **imperfect domains** (bandwidth quasi-cliques)
- Two stage iterative algorithm:
  1. estimate class matrices, fixing λ_i
  2. estimate λ_i, fixing class matrices
- E. Sefer, G. Duggal, and C. Kingsford. Deconvolution Of Ensemble Chromatin Interaction Data Reveals The Latent Mixing Structures In Cell Subpopulations, RECOMB 2015.

Sketch of how deconvolution works

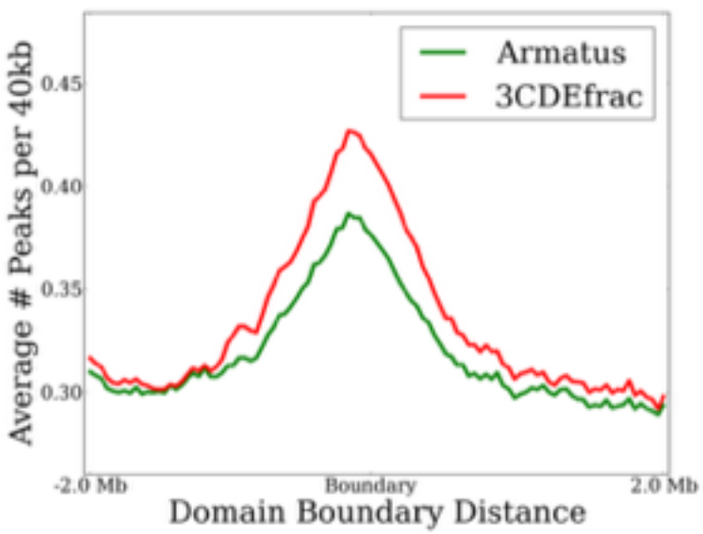
Bandwidth quasi-cliques:



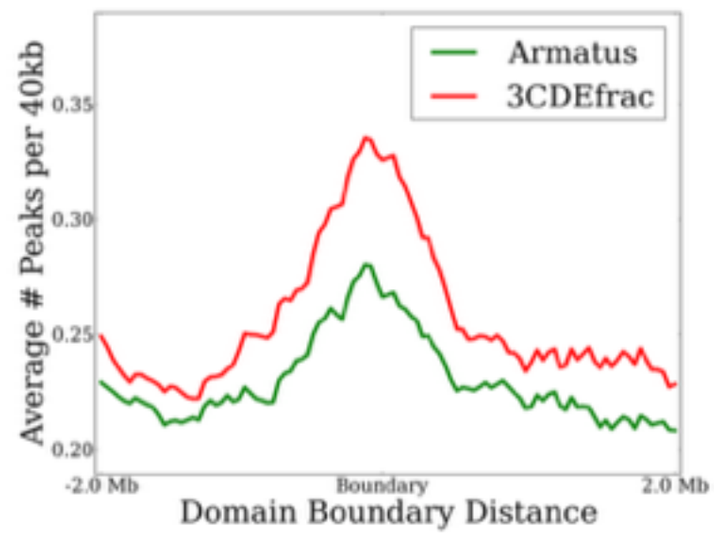
Iterative 2-step method for optimizing weights (X) & domains (Y):

1: $Y = \{(i,1) | \, i \in I\}$
2: **while** there is improvement in the objective (6) **do**
3: $\quad X = \text{argmin}_{A \in X,} \; Q(A, Y)$
4: $\quad Y = \text{argmin}_{B \in Y,} \; Q(X, B)$
5: **end while**

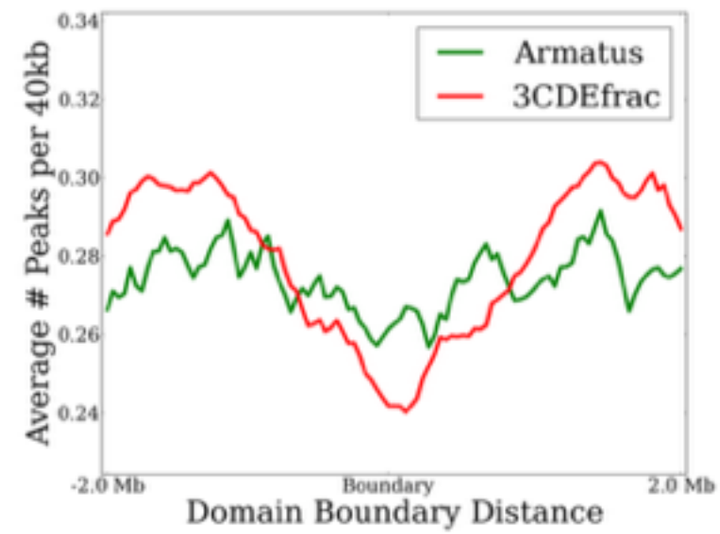
Deconvolution → Seemly better boundaries



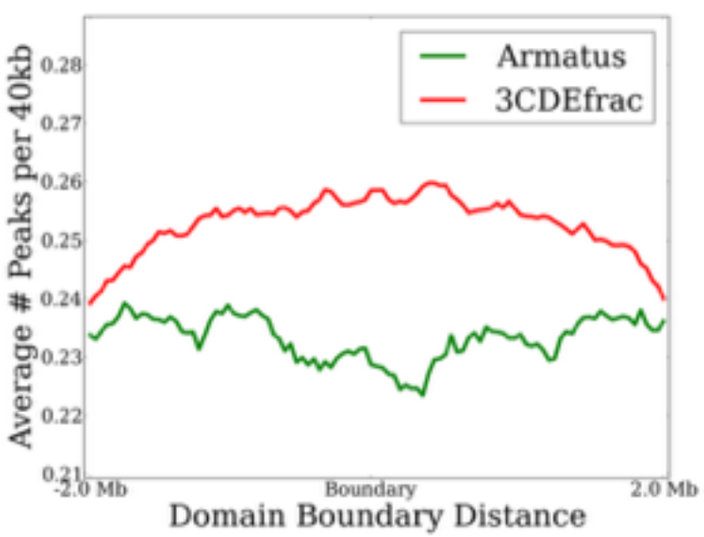
(a) H3K4me3 CD4$^+$



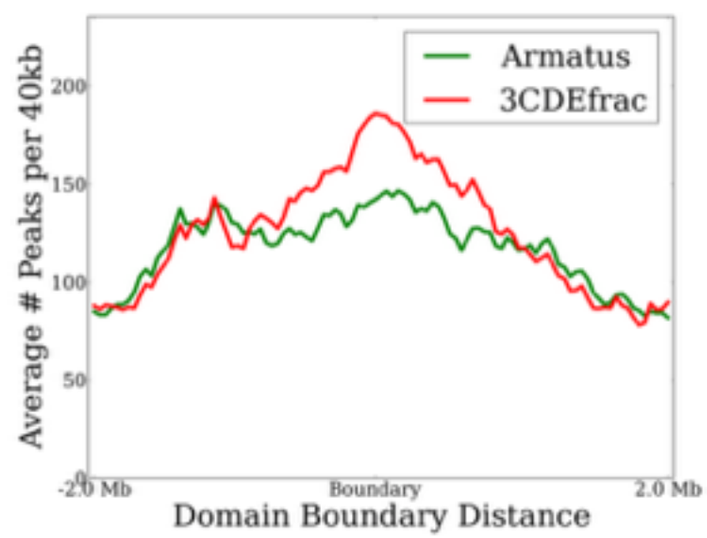
(b) H3K27ac CD4$^+$



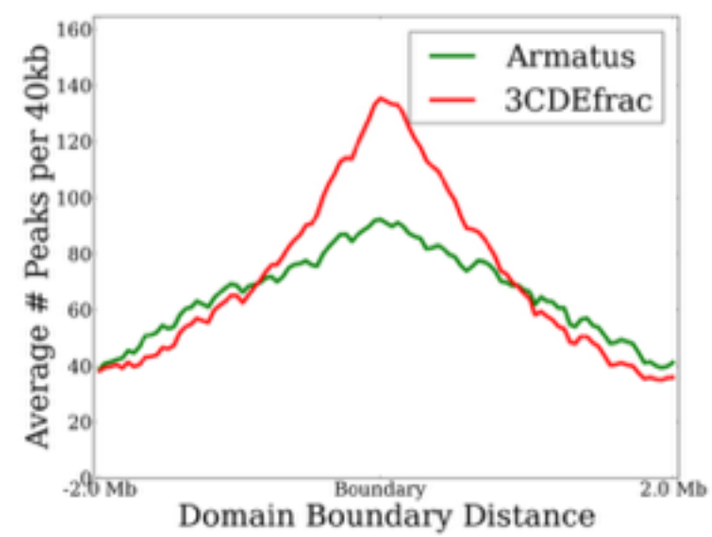
(c) H3K9me3 CD4$^+$



(d) H3K4me1 CD4$^+$



(e) H3K4me3 HeLa



(f) CTCF HeLa

Armatus:

- Identifies domains at multiple scales
- Diverse in size and location, better enrichment
- Requires a single parameter.
    - no assumptions about domain or boundary size, directionality, distribution of frequency values
- Fast: $O(n^2)$
    - IMR90 all chromosomes, all scales + consensus -- < 40 min on an 2.3Ghz Intel Core i5, 8Gb RAM (Java)
- Easily adapt block quality function $q(k, l, \gamma)$

Now: Working on methods to compare domains between cell types & species

Possible Renewal Contributions

- Relate spatial localization of transcription to (a) regulatory control, (b) phenotypes, (c) function more broadly [TR&D3]
  - May have some "structure-based" connection to [TR&D1]
- Tools for incorporating gene expression measurements into (a) pathway inference, (b) pathway evolution [TR&D2] (Sailfish/Salmon/SBT)
- Tools for comparing pathways and using pathway evolution to refine inferred pathways [TR&D2] (GHOST, PARANA1, PARANA2, NetArch, …)

Thanks

Darya Filippova

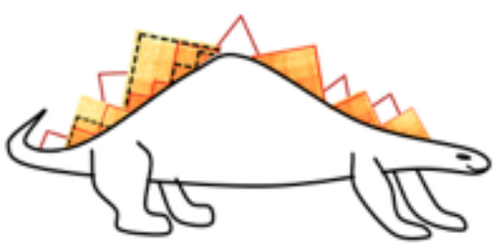
Rob Patro

Geet Duggal

Emre Sefer

Brad Solomon

Funding NIH R01 HG007104, R21 HG006913, T32 EB009403
NSF CCF-1256087, CCF-1319998
Sloan Research Fellow (C.K.)
Gordon and Betty Moore Foundation - Data Driven Discovery Investigator



www.cs.cmu.edu/~ckingsf/software/armatus