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

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

1gb to 20gb

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

Gene

Expression level

$\begin{align*}
\text{Expression level} & \quad \text{Gene} \\
100000 & \quad 0.01 \\
10000 & \quad 0.1 \\
5000 & \quad 1 \\
1000 & \quad 10 \\
200 & \quad 100 \\
65 & \quad 1000 \\
50 & \quad 10000 \\
10 & \quad 100000 \\
1 & \quad 1000000 \\
0.1 & \quad 10000000 \\
0.01 & \quad 100000000 \\
0 & \quad 1000000000 
\end{align*}$
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

A

<table>
<thead>
<tr>
<th></th>
<th>Alignment</th>
<th>Quantification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sailfish</td>
<td>2.27 h</td>
<td>10.19 h</td>
</tr>
<tr>
<td>RSEM</td>
<td>6.03 h</td>
<td>10.19 h</td>
</tr>
<tr>
<td>eXpress</td>
<td>6.03 h</td>
<td>10.19 h</td>
</tr>
<tr>
<td>Cufflinks</td>
<td>0.09 h</td>
<td>2.27 h</td>
</tr>
</tbody>
</table>

B

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
“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)
A prominent feature emerges from all four clusters: the arms are wound sinusoidally through space with roughly 1.5 period repeats per arm. The partial mirroring between clusters 1 and 2 and clusters 3 and 4 has the effect of causing the arms to be either intertwined (clusters 3 and 4) or separated (clusters 1 and 2). We favor the intertwined conformation, as the corresponding model clusters have lower variability (Figure S2) and lower IMP objective function scores (Table S2). However, it is possible that both conformations exist within a population of swarmer cells.

The parS Region Dictates the Orientation of the Entire Caulobacter Chromosome

Our models suggest that the parS sites play a direct role in organizing the swarmer cell chromosome. Such a finding is consistent with recent analyses that have suggested that these sequence elements are specifically anchored to the Caulobacter old cell pole through interactions with the ParB and PopZ proteins (Bowman et al., 2008; Ebersbach et al., 2008; Toro et al., 2008). Thus, we hypothesized that the orientation of the

![Diagram of genome spatial arrangement](image-url)
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)
“A” compartments = more open and loosely compacted

“B” compartments = more dense regions

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.


Distance is related to 1/frequency.

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

3C matrix

1Mbp, 2Mbp, 3Mbp
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)
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 persistent domains to form a single collection
How to find multiscale domains?

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

   \[
   \max \sum_{\text{domains}} q(\text{domain})
   \]

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

   \[
   \max \sum_{\text{domains at various scales}} p(\text{domain})
   \]

A - symmetric Hi-C matrix
Score dense blocks on the diagonal

block score (can be negative)

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

block weight

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

mean weight as a function of block size and resolution
Resolution parameter

block weight

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

\( \gamma = 0 : \) denominator becomes 1

\( \gamma = 1 : \) as used in [Goldberg 84]

\( \gamma = 2 : \) similar to weighted edge density

big domains

small domains
Resolution-Specific DP

\[
\text{OPT}_1'(l) = \max \left\{ \max_{k<l} \{ \text{OPT}_D(k-1) \}, \text{OPT}_D(l) \right\}
\]

\[
\text{OPT}_D(l) = \max_{k<l} \{ \text{OPT}_1'(k-1) + q'(k, l, \gamma) \},
\]

\[
q'(k, l, \gamma) = \begin{cases} 
q(k, l, \gamma) & \text{if } q(k, l, \gamma) > 0 \\
-\infty & \text{otherwise.}
\end{cases}
\]
Building a consensus of domains

domains = intervals, occurrence = weight

Weighted interval scheduling

\[ \text{OPT}_C = \max \begin{cases} 
\text{OPT}_C(j - 1) \\
\text{OPT}_C(c(j)) + p(a_j, b_j, \Gamma) 
\end{cases} \]

mark \( j' \) as non-domain
extend domain
Multiscale domains capture high frequency edges consistently.
Enrichment for structure-related genomic signals in the boundaries

CTCF
- transcriptional regulation
- insulator activity
- regulation of chromatin architecture

H3K27ac
- chromatin structure in eukaryotes
- form nucleosomes
- H3 most extensively modified

H3K4me3
- transcription activation/repression

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

[Dixon 2012]

[Image 1109x-13 to 1297x407]

Interactions upstream

[Image 1756x-111 to 1877x-88]

[Image 1774x-11 to 1871x407]

[Image 1167x-112 to 1250x-89]

[Image 1193x974 to 1214x1103]

[Image 1219x968 to 1833x1175]
Enrichment for chromatin marks

**CTCF in IMR90**

**CTCF in mESC**

**H3K27AC in mESC**

**H3K4me3 in mESC**

**Average # Peaks per 10Kbp**
More functional peaks in multiscale boundaries

<table>
<thead>
<tr>
<th>Signal</th>
<th>Boundaries (Dixon)</th>
<th>Boundaries (Armatus)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTCF (IMR90)</td>
<td>20%</td>
<td>44%</td>
</tr>
<tr>
<td>CTCF (mESC)</td>
<td>33%</td>
<td>72%</td>
</tr>
<tr>
<td>H3K4me3 (mESC)</td>
<td>30%</td>
<td>60%</td>
</tr>
<tr>
<td>H3K27ac (mESC)</td>
<td>23%</td>
<td>43%</td>
</tr>
</tbody>
</table>

Also: see peaks less often *within* multiscale domains
Analyses Enabled by High-quality Domains
Collect all optimal and near optimal-domains across scales into one set

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

Determine the percentage of all sufficiently different domain pairs $d_i, d_j$ where $d_i$ is completely contained within $d_j$ or vice-versa.

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

\[
\arg\max_D \log(P(D|W,H)) = \sum_{d=[s,e] \in \bar{D}} r_{se} x_{se} + \sum_{v \in V} E^e_v y_v
\]

\[
\bar{D} = \{ [s,e] | s,e \in V, e - s \geq 1 \}
\]

\[
r_{se} = E^b_s + E^b_e + \sum_{v=s+1}^{e-1} E^i_v
\]

- \( x \) 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

<table>
<thead>
<tr>
<th>Allowed modifications (human IMR90 to IMR90)</th>
<th>Coherence score (Normalized)</th>
</tr>
</thead>
<tbody>
<tr>
<td>28 histone modifications + Concave + Nonnegative *</td>
<td>1.00</td>
</tr>
<tr>
<td>28 histone modifications + Concave</td>
<td>0.99</td>
</tr>
<tr>
<td>28 histone modifications</td>
<td>0.97</td>
</tr>
<tr>
<td>H3K4me3, H3K79me2, H3K27ac, H3K9me3, H3K36me3, H4K20me1</td>
<td>0.94</td>
</tr>
<tr>
<td>H3K36me3, H3K4me1, H3K4me3, H3K9me3 + Concave + Nonnegative</td>
<td>0.94</td>
</tr>
<tr>
<td>H3K36me3, H3K4me1, H3K4me3, H3K9me3 + Concave</td>
<td>0.93</td>
</tr>
<tr>
<td>H3K36me3, H3K4me1, H3K4me3, H3K9me3</td>
<td>0.92</td>
</tr>
</tbody>
</table>
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 $\lambda_i$
  2. estimate $\lambda_i$, fixing class matrices

Sketch of how deconvolution works

Bandwidth quasi-cliques:

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

1: \( Y = \{(i, 1) \mid i \in I\} \)
2: while there is improvement in the objective (6) do
3: \( X = \operatorname{arg\min}_{A \in X} Q(A, Y) \)
4: \( Y = \operatorname{arg\min}_{B \in Y} Q(X, B) \)
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

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