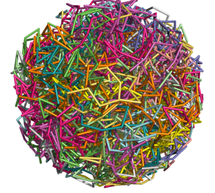
HYPOTHESIS TESTING OF 3D CO-LOCALIZATION OF GENOMIC ELEMENTS

Jonas Paulsen[1], Tonje G. Lien[2], Geir Kjetil Sandve[2], Lars Holden[3], Ørnulf Borgan[2], Ingrid K. Glad[2] and Eivind Hovig[1,2]

[1] Oslo University Hospital, Norway, [2] University of Oslo, Norway, [3] Statistics For Innovation, Norwegian Computing Center, Norway



ABSTRACT

The three-dimensional (3D) structure of chromatin is crucial for fundamental processes in the nucleus. We are interested in evaluating whether certain regions in the genome are spatially closer to each other than what would be expected by chance. This problem has recently been studied by Witten et al. 2012. In Paulsen et al. 2013 we addressed the important issue of dependencies between interaction frequencies in 3D datasets when estimating p-values.

Hi-C Data

Let genomic element a_i be the element that starts on base pair i on chromosome a. Let $X_{a_i b_j}$ be the **interaction frequency** between genomic elements a_i and b_j, the number of times the Hi-C method detects that a_i and b_j are spatially close (X is normalized and do not need to be an integer).

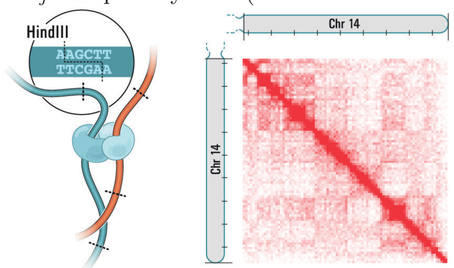


Figure 1: Cross-linking the DNA (left) and the use of next-generation sequencing determine the interaction frequency $X_{a_i b_j}$. The matrix (right) of one arm on chromosome 14, show all possible interaction frequencies $X_{14_i,14_j}$. Both figures by Lieberman-Aiden et al. 2009

BIG DATA: Using resolution (bin size 100k base pairs) on the human genome with about 3.2 billions base pairs, gives us 32000 bins and 500 millions possible interactions.

Highly dependent data

The following dependencies are taken into account in the hypothesis test:

- The expectation $E(X|\delta)$ and standard deviation $sd(X|\delta)$ are dependent on δ (see Figure 2, 3).
- The dependency between pairs of X are high if the genomic elements are sequential close along the genome (Transitive relation).
- X is highly dependent on the GC-content (Lieberman-Aiden et al. 2009) and the sequential positioning of its genomic elements along the chromosome (Imakaev et al. 2012).

$$\delta = \begin{cases} |j-i| & \text{if } a = b \\ \infty & \text{if } a \neq b. \end{cases}$$

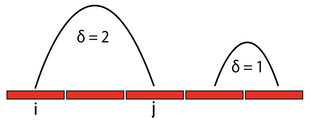


Figure 2: The sequential distance corresponding to an interaction.

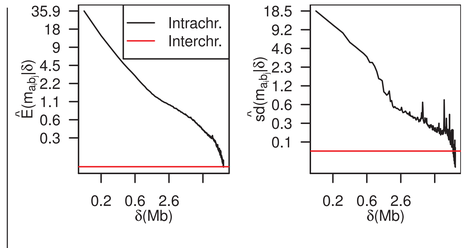


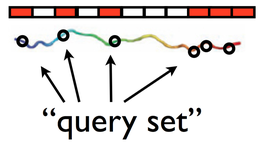
Figure 3: The sample mean $\hat{E}(X|\delta)$ and sample standard deviation $\hat{sd}(X|\delta)$ of an interaction frequency X.

Statistical method

Hypotheses:

H_0: The query set has the same 3D co-localization as expected by chance,

H_1: The query set has more 3D co-localization than expected by chance.



Let the **test statistic** measure the amount of 3D co-localization in a query set Q, where M is the number of interactions in the sum, as the following

$$t = \frac{1}{M} \sum_{a_i, b_j \in Q} \frac{x_{a_i b_j} - \hat{E}(X|\delta)}{\hat{sd}(X|\delta)}.$$

For each chromosome a, calculate d_a the set of all sequential distances between all pairs of consecutive genomic elements in the query set. Repeat the following **randomization procedure** R times (for $r \in 1 \cdots R$):

- For each chromosome a, permute the sequential position along the genome of the genomic elements of interest, while preserving d_a.
- Let t_r be the test statistic based on the random set.

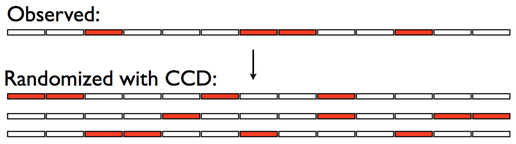


Figure 4: Toy example of the permutation.

Calculate the **p-value**:

$$p = \frac{\sum_{r=1}^{R} I(t_r \geq t_{\text{obs}}) + 1}{R + 1}$$

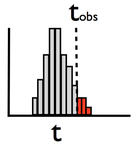


Figure 5: The estimated p-value is the number of resamples test statistics larger than t_{obs} (based on the observed query set of interest)

Additionally, with **the domain randomization** we can maintain structural properties, such as the GC-content or the relative positioning along chromosome arms, by permuting within predefined domains.

Uniformly distributed p-values under H_0

Based on random walk we simulate intra- and inter-chromosomal interaction frequencies, and our method correctly gives uniformly distributed p-values for every considered query set under H_0.

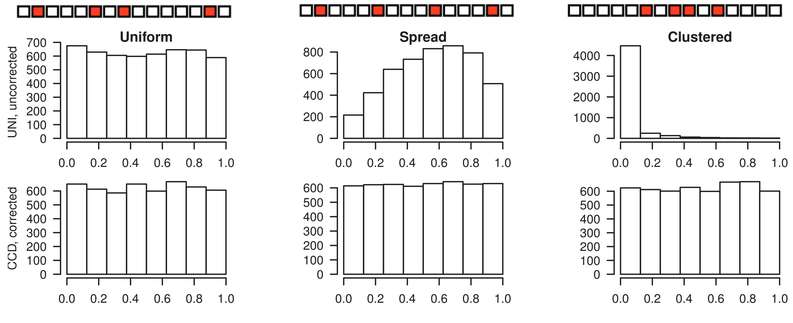
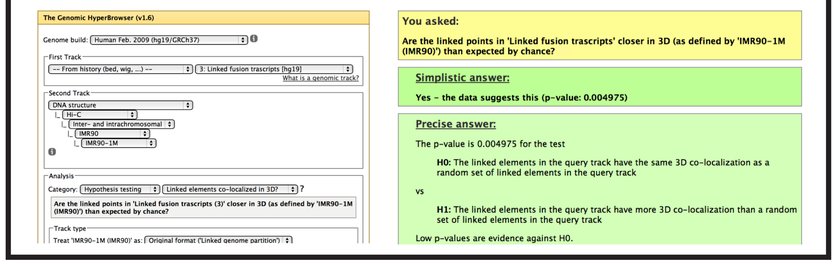


Figure 6: We compared our method (bottom row) to a simpler method using uniformly resampled sets and the mean of $X_{a_i b_j}$ as test statistic (upper row). The different columns represent different configurations of the query set.

Software

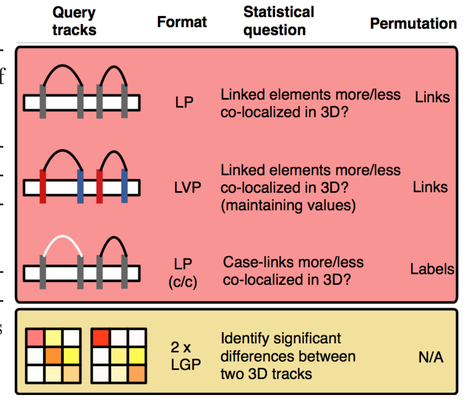
All algorithms have been implemented in a publicly available statistical web toolkit called HiBrowse, at *hyperbrowser.uio.no/3d*



Expanded testing

In the following up manuscript Paulsen, J. et al. 2014 (in press) we expand the number of possible hypothesis tests. The user:

- specifies particular interactions between a set of genomic elements
- marks a sub set of interactions as cases, and compare those to a set of control interactions
- finds all statistically significant differences between two Hi-C data sets (based on edgeR by Robinson et al., 2010).



Summary of Paulsen et al. 2013

We find strong dependency in interaction frequencies between contacts with low sequence-based distance which strongly affect the p-value estimation. To obtain valid and biologically meaningful p-value, it is essential to take such dependencies into account in the resampling steps. In addition we handle intra- and inter-chromosomal interactions both separately and jointly. The results are presented with p-values and enrichment scores. All software is available online at *hyperbrowser.uio.no/3d* .

References

Witten, D. M. et al. (2012) *Nucleic Acids Res.*, **40**, 3849–3855 .
Lieberman-Aiden, E. et al. (2009) *Science*, **326**, 289–293.
Imakaev, M. et al. (2012) *Nat Methods.*, **9**, 999–1003.
Sandve, G. et al. (2010) *Genome biol.*, **11**, R121.
Paulsen, J. et al. (2013) *NAR*, **41** 5164–5174
Paulsen, J. et al. (2014) *Bioinformatics*, **In press**