nature structural & molecular biology

Article

https://doi.org/10.1038/s41594-022-00892-7

Diverse silent chromatin states modulate genome compartmentalization and loop extrusion barriers

Supplemental Note

Spectral clustering of Hi-C data

Here we describe a spectral clustering algorithm to identify loci with common long-range contact frequency profiles in Hi-C data. As in previous work, we focus on interchromosomal (*trans*) contact frequencies. Because *trans* contacts are experimentally sampled less frequently than intrachromosomal (*cis*) contacts, this strategy works best on deeply sequenced contact maps with strong pattern contrast in *trans*. However, it provides the advantages of not having to detrend intra-arm and inter-arm polymer scaling relationships and of working on more than one chromosomal arm at a time. For example, [1] used contact frequencies between two sets of disjoint chromosomes (odd and even-numbered) and clustered *trans* Hi-C data along each axis (rows and columns) separately at 100-kb resolution, followed by harmonizing cluster identities between the two axes.

Clustering the leading eigenvectors of a suitably normalized contact matrix (i.e., spectral clustering) provides a scalable alternative to clustering the complete matrix directly that also dampens noise and highlights cluster structure in the data [2]. In [3], this was done separately on each chromosome using *cis* data, followed by a recursive clustering step using *trans* data to harmonize cluster identities between chromosomes. In this study we apply dimensionality reduction using global eigendecomposition on *trans* contact frequencies from genome-wide balanced 50kb-resolution Hi-C maps. This provides a single collection of eigenvectors for clustering, requiring no harmonization between chromosome arms or sets thereof. As an additional benefit, it also facilitates further dimensionality reduction (embedding) and visualization.

Before eigendecomposition, because *trans* translocations contaminate interaction profiles and will exert an influence on higher order eigenvectors, we excluded such regions from analysis. We manually identified and excluded three large translocated segments in HCT116 (chr8: 67.35 Mb – end; chr16: 78.93 Mb – end; chr17: 43.40 Mb – end) based on published karyotype analysis [4] narrowed down by visual inspection of Hi-C data in HiGlass [5]. Unfortunately, structural variations in DKO were too widespread to systematically exclude, so DKO clustering results were omitted from this study.

To mask the influence of *cis* data, we followed the same procedure described in [6], where *cis* pixels in the contact matrix are replaced with randomly sampled *trans* pixels from the same row or column. The resulting matrix was then re-balanced. In order to standardize the spectrum of eigenvalues, the matrix was scaled such that rows and columns sum to 1 and eigendecomposition was done without the prior mean-centering step used in [6]. The decomposition of this $n \times n$ normalized affinity matrix A (interpreted as a weighted graph) can be expressed as a sum of rank-1 outer products:

$$A = \lambda_0 \vec{e}_0 \vec{e}_0^T + \lambda_1 \vec{e}_1 \vec{e}_1^T + \lambda_2 \vec{e}_2 \vec{e}_2^T + \ldots + \lambda_{n-1} \vec{e}_{n-1} \vec{e}_{n-1}^T, \tag{1}$$

where $\lambda_0 = 1$, $\vec{e_0} = \frac{1}{\sqrt{n}}\mathbb{1}$, $||\vec{e_i}||= 1$ and

$$1 \geq |\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_{n-1}| \geq 0$$

The largest eigenvalue λ_0 is 1 and its multiplicity depends on the number of connected components of the graph whose edge weights are given by A. Because a Hi-C matrix of sufficient coverage represents a single connected component (ignoring filtered bins), λ_0 is uniquely equal to 1 and is associated with a constant-valued eigenvector $\vec{e}_0$ that may be discarded for clustering purposes [7, 8]. The remaining eigenvalues lie in the range $(-1, 1)$ and those with relatively large modulus are most informative for clustering [2, 9]. Interestingly, the next leading eigenvector ($\vec{e}_1$) is identical to the usual A/B compartment eigenvector (from the centered version of the same balanced matrix) and also happens to be the well-known "Fiedler vector" in spectral clustering [8]. Therefore, when A is balanced, the classical bipartitioning of genomic loci based on the first eigenvector of centered A (or equivalently, the first principal component from performing PCA on A) happens to be a special case of spectral clustering. Even when A is not balanced, these procedures are closely related.

To extend this approach in order to obtain a finer and more informative partition, we selected the m leading eigenvectors after $\vec{e}_0$ ($m < n-1$). The leading eigenvalues and associated eigenvectors of A were calculated using the `eigsh` routine from *numpy* [10], in descending order of eigenvalue modulus (i.e. not respecting algebraic sign). The approximation can be written as

$$
\begin{aligned}
A &\approx \vec{e}_0\vec{e}_0^T + \lambda_1\vec{e}_1\vec{e}_1^T + \ldots + \lambda_m\vec{e}_m\vec{e}_m^T \\
&= \vec{e}_0\vec{e}_0^T + \\
&\quad \operatorname{sign}(\lambda_1)\sqrt{|\lambda_1|}\vec{e}_1 \cdot (\sqrt{|\lambda_1|}\vec{e}_1)^T + \ldots + \\
&\quad \operatorname{sign}(\lambda_m)\sqrt{|\lambda_m|}\vec{e}_m \cdot (\sqrt{|\lambda_m|}\vec{e}_m)^T .
\end{aligned} \tag{2}
$$

Motivated by equation (2), the individual unit-normed eigenvectors (excluding $\vec{e}_0$) were weighted as shown ($\sqrt{|\lambda_1|}\vec{e}_1$, $\sqrt{|\lambda_2|}\vec{e}_2$, ..., $\sqrt{|\lambda_m|}\vec{e}_m$), concatenated as columns, and k-means clustering was applied to the rows using *scikit-learn*'s `KMeans` estimator [11]. The estimator was run with 100 centroid initializations using the `kmeans++` initializer, which returned the best run in terms of inertia. As a heuristic, m was chosen to be the maximum number of eigenvalues before the first negative eigenvalue appeared (i.e., before succeeding eigenvalues began oscillating around 0), though empirically for a given k, clustering was observed to be stable beyond that number as long as genomic regions undergoing translocations were excluded from analysis. We produced cluster assignments for a range of k for GM12878 [1] and both unsynchronized untreated and 6h Auxin-treated Rad21-AID HCT116 maps [12], calculated silhouette scores and visually compared cluster profiles to a large number of independent genomic tracks. The final number of clusters for HCT116 ($k = 8$) was chosen based on a balance of clustering metrics and interpretability.

Finally, we note that readers familiar with spectral partitioning of graphs may be more accustomed to discussions of the eigenvectors of the graph Laplacian matrix L rather than the affinity matrix A. The Laplacian is $L = D - A$, where D is a diagonal matrix with the node degrees $d_1, \ldots, d_n$ on the diagonal. The most common normalized transformations of the Laplacian used for spectral clustering are $L_{sym} = D^{-1/2}LD^{-1/2}$ and $L_{rw} = D^{-1}L$ [8]. Note that in our case, A is a stochastic matrix (i.e. a graph with constant degree 1) and its (normalized) Laplacian is simply $I - A$, which has the exact same eigenvectors as A and shifted eigenvalues $\lambda_i' = 1 - \lambda_i$.

References

[1] S. S. Rao, M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander, *et al.*, "A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping," *Cell*, vol. 159, no. 7, pp. 1665–1680, 2014.

[2] M. Brand and K. Huang, "A unifying theorem for spectral embedding and clustering," in *International Workshop on Artificial Intelligence and Statistics*, pp. 41–48, PMLR, 2003.

[3] B. Lucic, H.-C. Chen, M. Kuzman, E. Zorita, J. Wegner, V. Minneker, W. Wang, R. Fronza, S. Laufs, M. Schmidt, *et al.*, "Spatially clustered loci with multiple enhancers are frequent targets of HIV-1 integration," *Nature communications*, vol. 10, no. 1, pp. 1–12, 2019.

[4] S. Langer, J. B. Geigl, S. Ehnle, R. Gangnus, and M. R. Speicher, "Live cell catapulting and recultivation does not change the karyotype of HCT116 tumor cells," *Cancer genetics and cytogenetics*, vol. 161, no. 2, pp. 174–177, 2005.

[5] P. Kerpedjiev, N. Abdennur, F. Lekschas, C. McCallum, K. Dinkla, H. Strobelt, J. M. Luber, S. B. Ouellette, A. Azhir, N. Kumar, *et al.*, "HiGlass: web-based visual exploration and analysis of genome interaction maps," *Genome biology*, vol. 19, no. 1, pp. 1–12, 2018.

[6] M. Imakaev, G. Fudenberg, R. P. McCord, N. Naumova, A. Goloborodko, B. R. Lajoie, J. Dekker, and L. A. Mirny, "Iterative correction of Hi-C data reveals hallmarks of chromosome organization," *Nature methods*, vol. 9, no. 10, pp. 999–1003, 2012.

[7] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in neural information processing systems*, pp. 849–856, 2002.

[8] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.

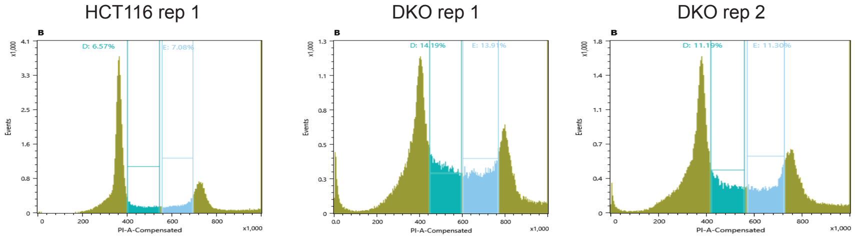
[9] S. Franzini, M. Di Stefano, and C. Micheletti, "essHi-C: essential component analysis of Hi-C matrices," *Bioinformatics*, 02 2021. btab062.

[10] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, "Array programming with NumPy," *Nature*, vol. 585, pp. 357–362, Sept. 2020.

[11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[12] S. S. Rao, S.-C. Huang, B. G. St Hilaire, J. M. Engreitz, E. M. Perez, K.-R. Kieffer-Kwon, A. L. Sanborn, S. E. Johnstone, G. D. Bascom, I. D. Bochkov, *et al.*, "Cohesin loss eliminates all loop domains," *Cell*, vol. 171, no. 2, pp. 305–320, 2017.

Supplementary Figure 1



FACS sorting for Repli-seq based on propidium iodide staining.
Early vs late fraction gating strategy indicated.