Dimensionality Reduction and Clustering Techniques for Hi-C Genomic Data Analysis

Benjamin Jagdeo [1], Kini Chen [2], Anna Rasheed [3], Saladin Shaurov [4], Arian Hashemzadeh [4], Elena Tuzhilina [4]
[1] University of Guelph, [2] McGill University, [3] McMaster University, [4] University of Toronto

Abstract

Obtaining the 3D structure of a DNA strand within a cell can enable researchers studying genetics to analyze the structural and functional properties of the DNA and cell. However, it is extremely difficult to read this 3D structure using current methodologies. Hi-C analysis serves as a tool to study this structure by measuring the contact frequencies between different genomic loci in the strand. Unfortunately, Hi-C data is extremely noisy and high-dimensional, which presents numerous challenges in identifying structural properties and conducting inference. In this project, we aim first to reduce the noise through the preprocessing stage by reducing our dataset. Then, we aim to visualize the global structure and properties of our data through various dimensionality reduction methods such as PCA, UMAP, and t-SNE. The goal of the project is to evaluate the performance of these approaches in effectively separating the 4 different cell-types available in our dataset through the projection of our data onto a reduced dimensional subspace.

Objectives

We aim to classify different human cell types (GM12878, HAP1, HeLa, K562) using chromosomal contact data derived from the Hi-C analysis. This project explores techniques in high-dimensional data analysis, including dimensionality reduction methods for visualization in lower dimensional subspaces, clustering for classification, parameter tuning, and performance evaluation using numerical metrics. By analyzing contact variability through high-dimensional matrices and selecting statistically significant features, we aim to uncover the main patterns that differentiate between cell types.

Methods

Principle Component Analysis (PCA) is a dimension reduction technique used to analyze the properties of high-dimensional data. Assuming the centered data $\bar{X}$ lies on an *r*-dimensional affine subspace in $\mathbb{R}^d$, PCA aims to find an orthogonal basis for this subspace by finding the eigenvectors v of the sample covariance matrix S. By projecting the original data onto the top *r* eigenvectors stored in V_r, we retain the maximum proportion of the original data's variance ($\sum_{i=1}^{d} \lambda_i$) while compressing the data to Z in *r*-dimensional space.

$$\mathbf{S = (1/n)\,\bar{X}^T\,\bar{X}} \qquad \mathbf{S v_i = \lambda_i v_i} \qquad \mathbf{Z = \bar{X} V_r}$$

Kernel Principle Component Analysis (KPCA) is an extension of PCA which allows for nonlinear dimensionality reduction by first mapping the data into a higher-dimensional space using a kernel function and then performing PCA in that space — enabling it to capture complex, curved structures that standard PCA cannot. Typically, a nontrivial arbitrary function $\mathbf{\Phi}: \mathbb{R}^d \rightarrow \mathbb{R}^N$ is chosen, and the kernel is defined as the inner product space $\mathbf{K = k(x,y) = (\Phi(x), \Phi(y)) = \Phi(x)^T \Phi(y)}$. The function $\mathbf{\Phi}$ can map points to an arbitrarily high-dimensional space which allows the data to be easily linearly separable, and the kernel allows PCA to be conducted without ever calculating the values of the function $\mathbf{\Phi}$**.** We apply a simpler version of KPCA which maps the contact data to vectors of cell summary-statistics and conducts PCA as usual in this new space. We do not use the kernel in conducting PCA due to the explicit calculation of $\mathbf{\Phi}$**,** however, the mapping allows us to handle non-linearity in the contact data.

Uniform Manifold Approximation and Projection (UMAP) is a nonlinear dimension reduction technique designed to preserve both local and global properties of high-dimensional data. It assumes the data lies uniformly on a locally connected Riemannian manifold with a locally constant metric defined. UMAP constructs a weighted k-nearest neighbor graph in the original space and finds a low-dimensional embedding by optimizing the similarity between this graph and one in the reduced space. This method often outperforms PCA in capturing the global structure of complex datasets, as it can capture nonlinear structures in the data more effectively.

t-distributed Stochastic Neighbor Embedding (t-SNE) is another dimension reduction method to visualize high-dimensional data onto lower dimensional subspaces. The algorithm calculates the pairwise distances and variances (similarities) between data points that follows a Gaussian distribution to compute the probability a datapoint being the neighbor of another. Then it creates an initial set of low-dimensional counterpoints of the data based on a similarity matrix which follows t-distribution. The data is updated to minimize the Kullback–Leibler (KL) divergence so that low-dimensional data approximates the distribution of high-dimensional data.

Preprocessing

Raw Hi-C data was first preprocessed to form block-diagonal contact matrices where each block represents intra-chromosomal contacts between gene loci pairs for that chromosome.

For, we chose to simplify the contact frequency data and emphasize the presence or absence of interactions by binarizing the contact matrices. Specifically, any non-zero contact count was converted to 1, while zero counts remained unchanged.

The upper triangular part of each cell's block-diagonal contact matrix was flattened into a one-dimensional feature vector and stacked vertically to form the data matrix X.

Number of features $= \sum_{c \in \mathcal{C}} \frac{n_c(n_c - 1)}{2}$, where C is the set of chromosomes we kept, and n_c is the number of loci in C.

Our processed data set, X, is of dimension 1958 × 236562 before filtering out the least informative chromosomes, and after filtering, X* reduces to dimension 1958 × 139448.

To reduce the noise, filtering was applied at two levels. First, we ranked cells by the mean number of non-zero entries in their contact matrices and removed the bottom 25 percentile of cells (sparse). Then, we ranked chromosomes based on their average absolute PC1 loading values and retained the top 75% of chromosomes for further analysis.

Results

PCA

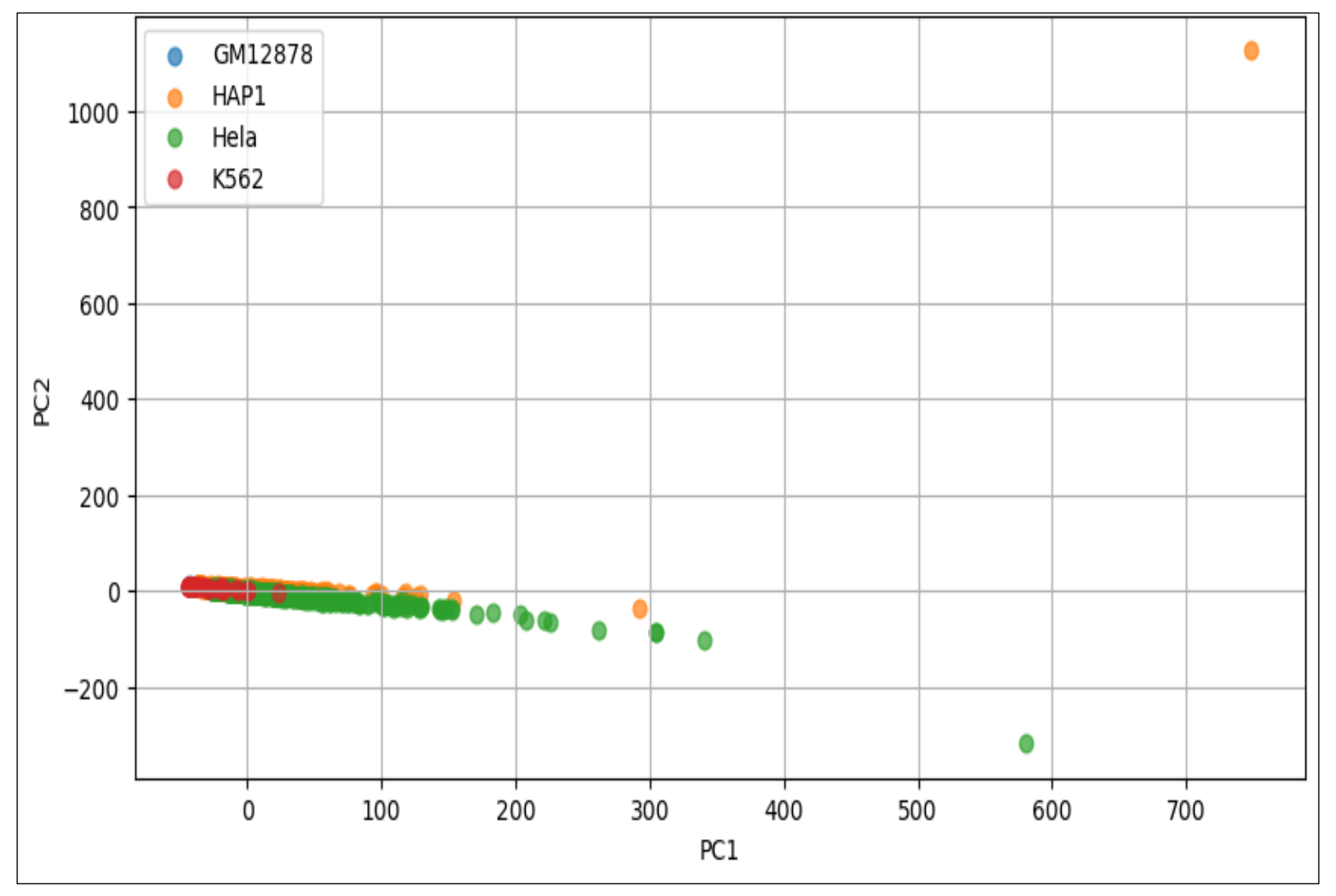

Fig. 1: PCA projection of the filtered data, showing significant overlap in cell types and serving as a baseline method

PCA + UMAP

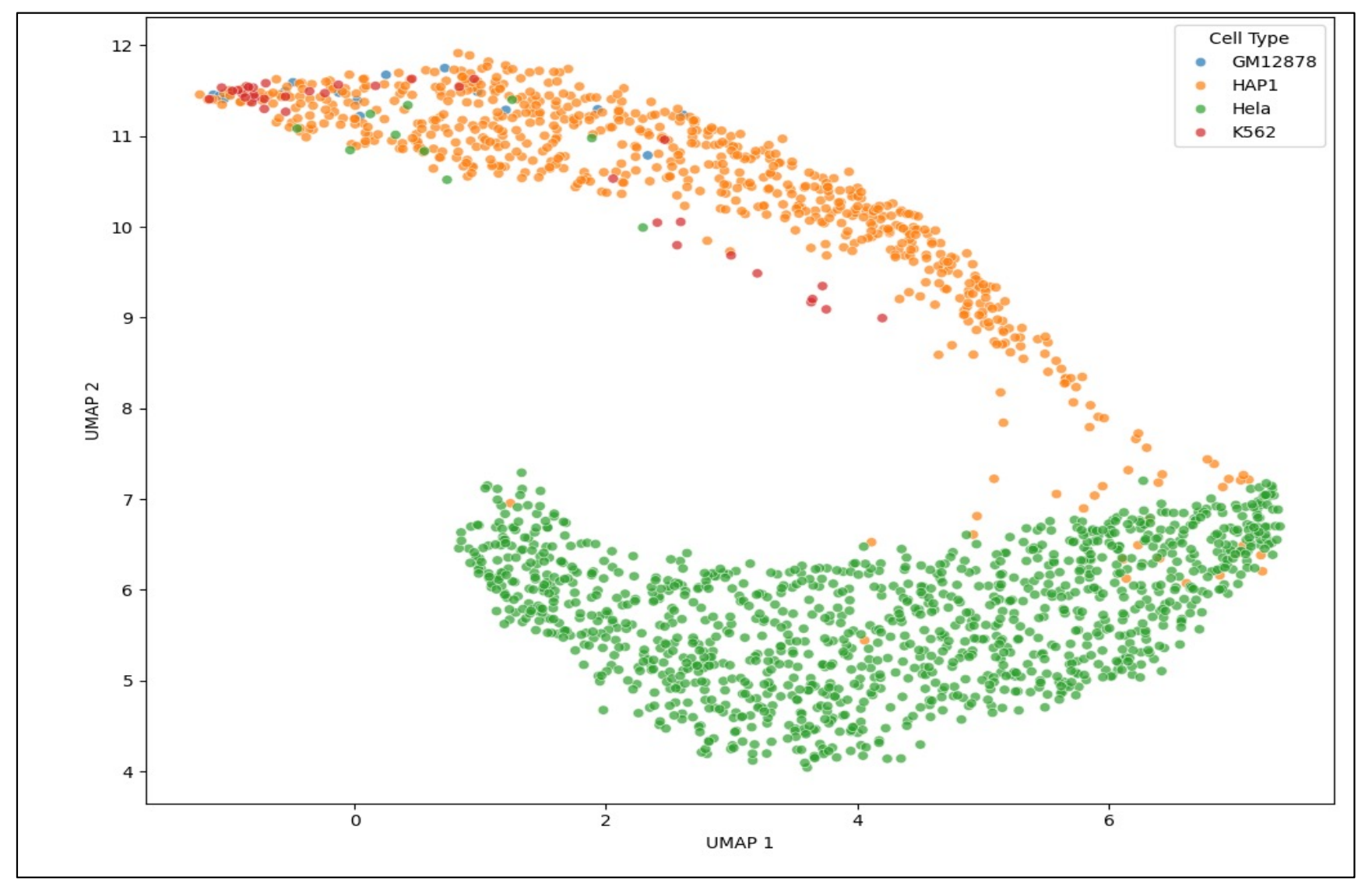

Fig. 2: UMAP projection of the filtered data following PCA-based dimensionality reduction.

Weighted PCA + UMAP

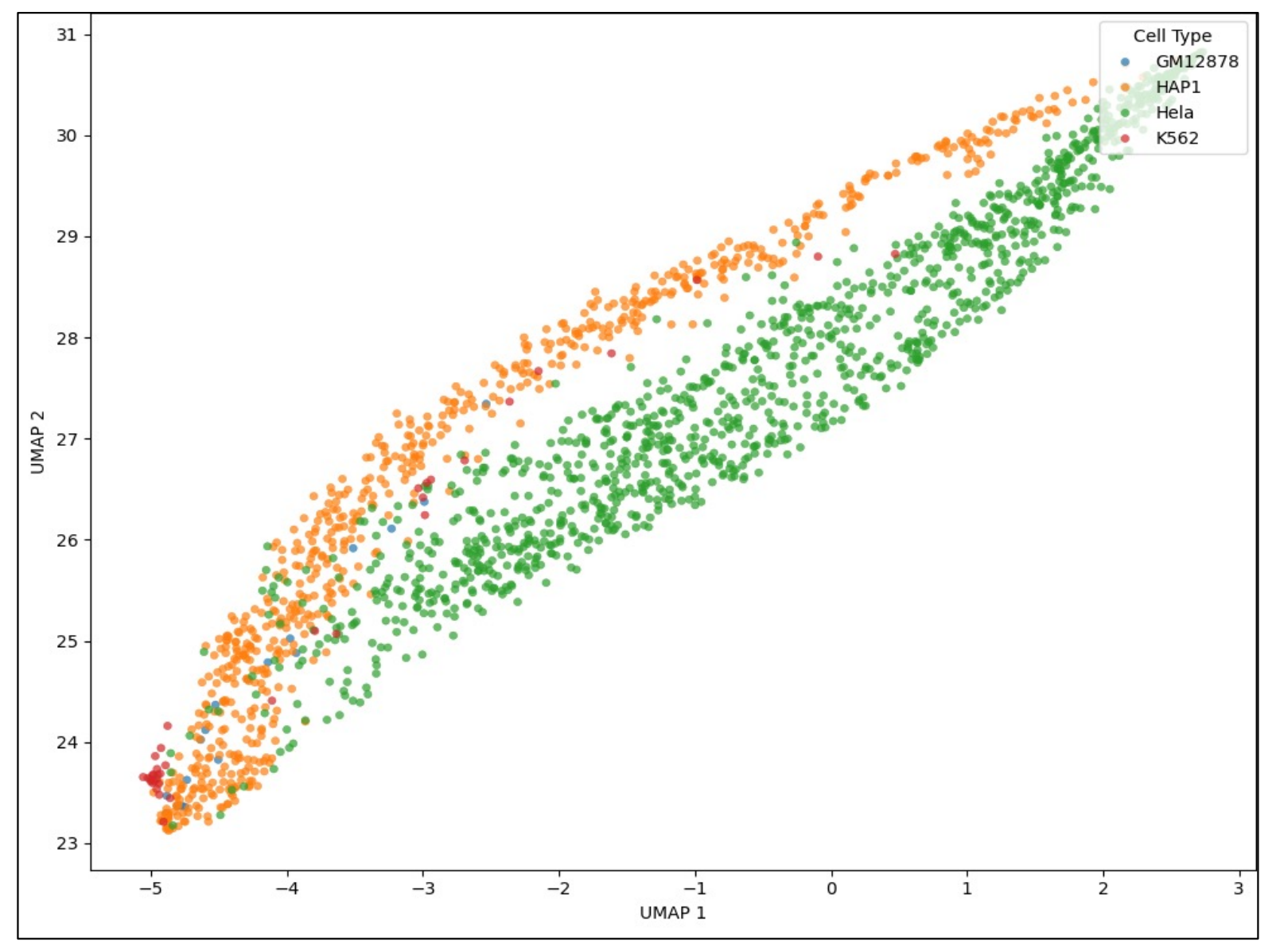

Fig. 3: UMAP projection after applying feature weights prior to PCA on the filtered data

t-SNE

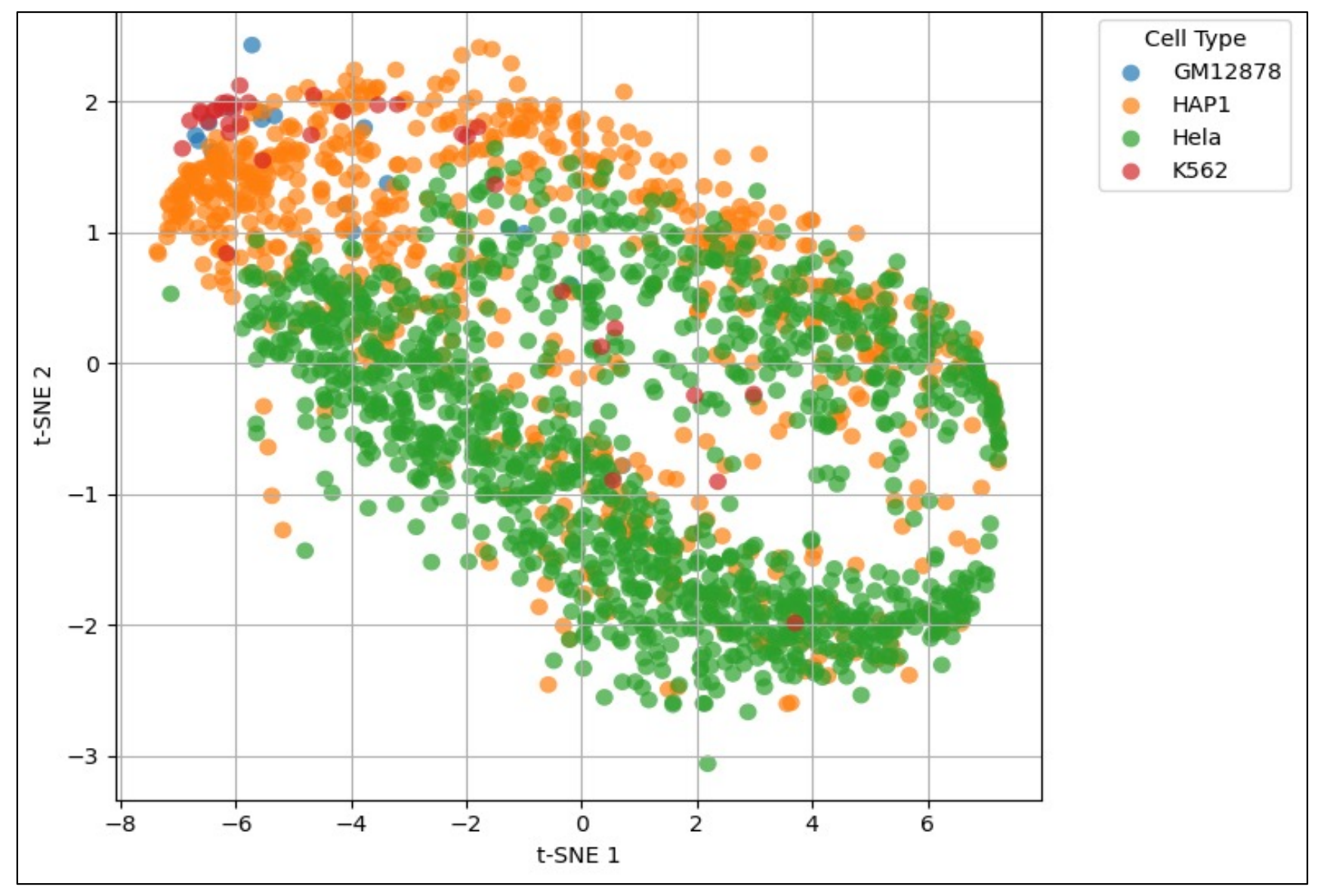

Fig. 4: t-SNE applied to the summary statistics data as an additional method

Kernel PCA

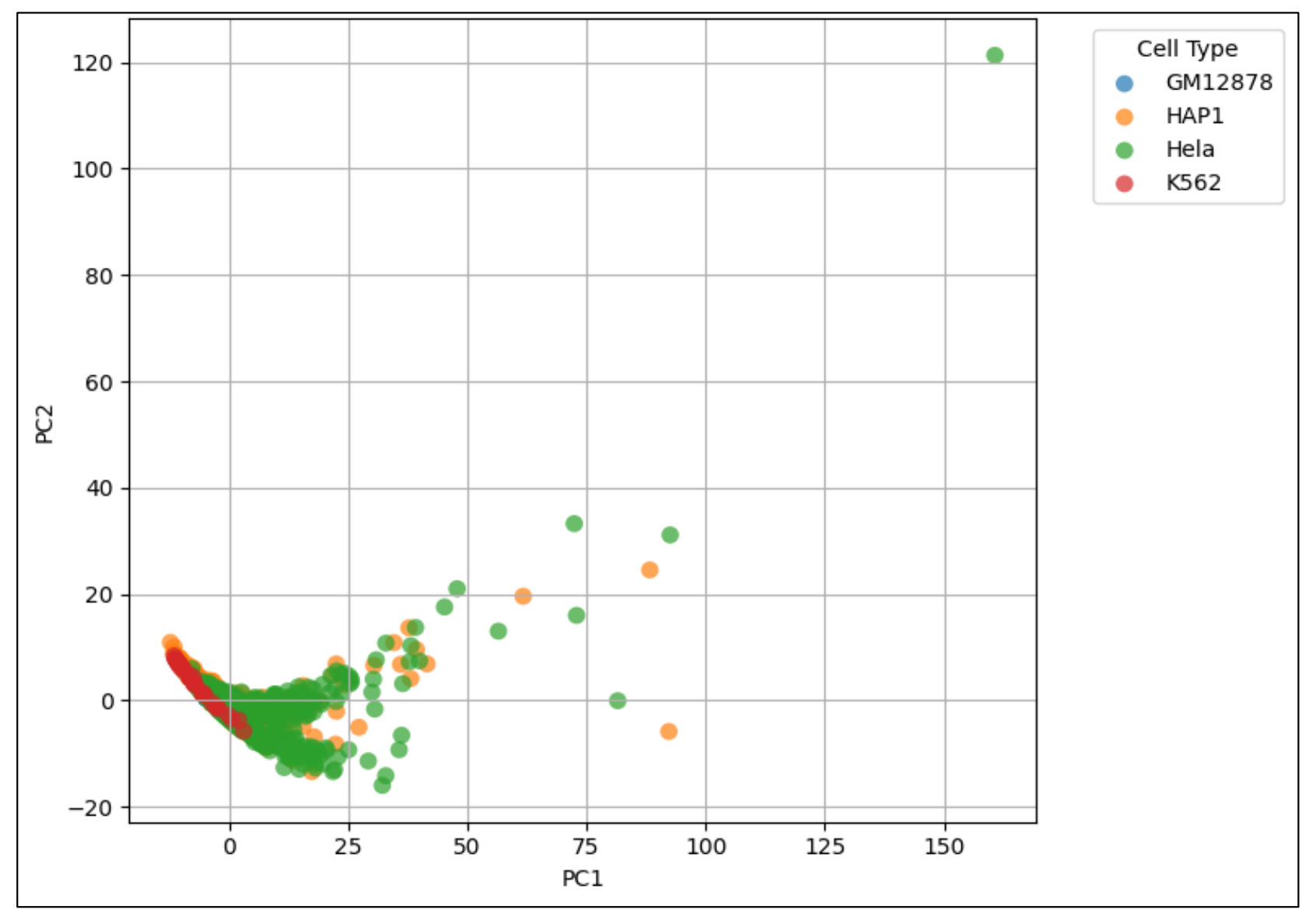

Fig. 5: PCA on summary statistics. Contact matrices were nonlinearly mapped to summary-statistics space prior to PCA.

Kernel PCA + UMAP

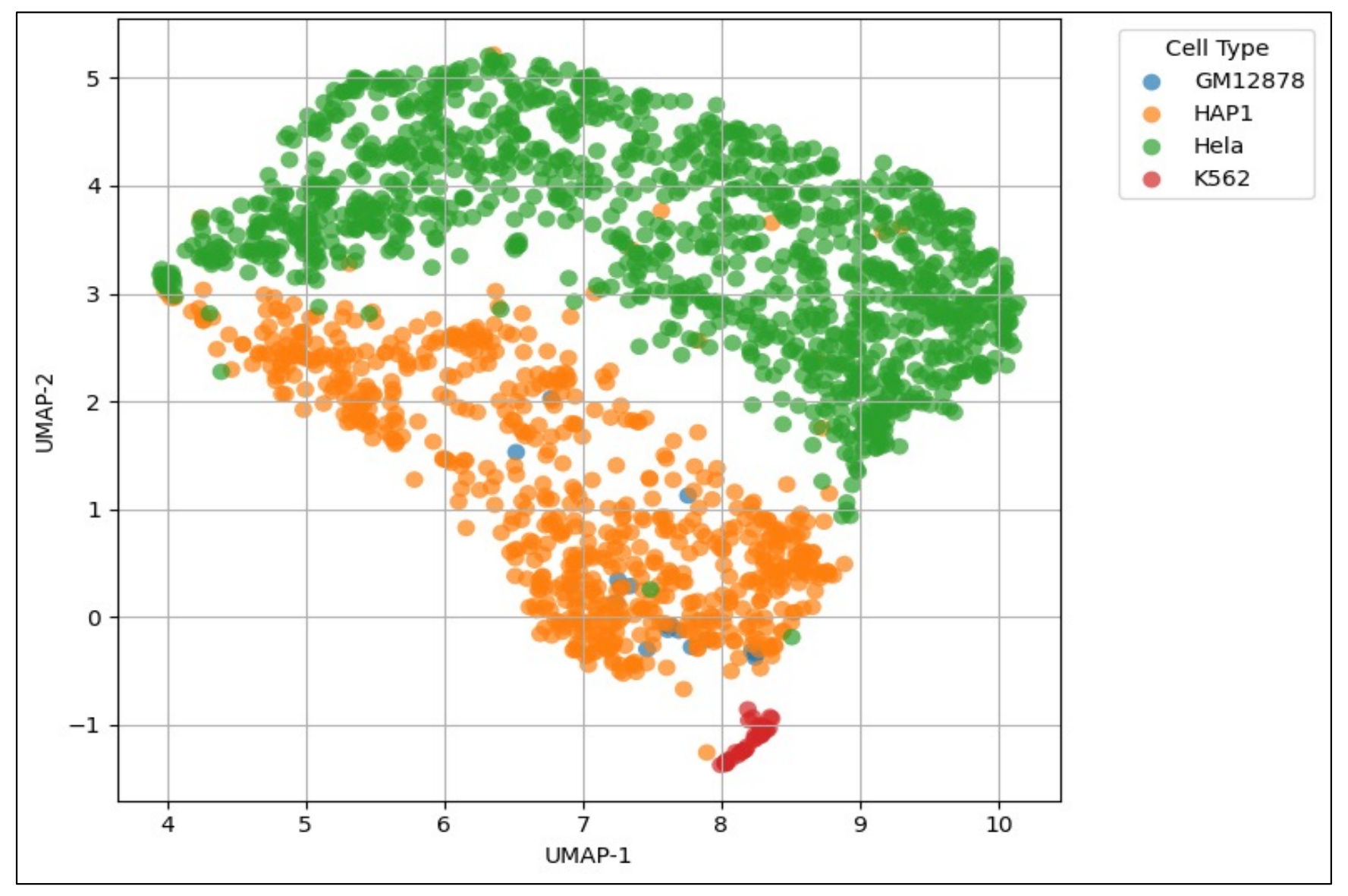

Fig. 6: UMAP projection of the summary statistics data, following PCA-based dimensionality reduction

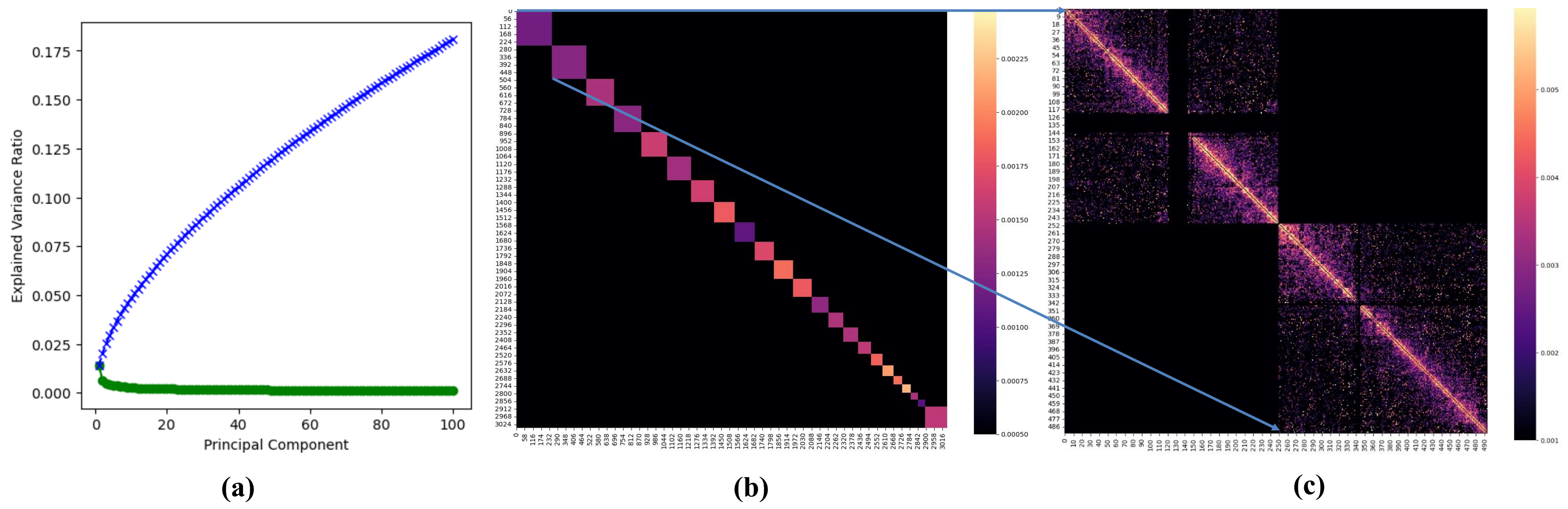

Fig. 7: **(a)** Scree plot showing the individual and cumulative variance explained by each principal component after filtering chromosomes. The first few components only capture a minimal amount of variance. **(b)** Heatmap of PC1 loading values per gene loci pair averaged across chromosomes. Higher (brighter) blocks indicate chromosomes that explain more variance between cells. **(c)** Submatrix of PC1 corresponding to chromosomes 1 and 2. Gene loci pairs closer in distance appear to be more informative.

	K-means (ARI)		Spectral Score (ARI)		Run Time (s)	
	Unfiltered	**Filtered**	**Unfiltered**	**Filtered**	**Unfiltered**	**Filtered**
PCA	0.07	0.09	0.12	0.12	158.95	80.18
PCA + UMAP	0.45	0.42	0.44	0.42	325.50	120.20
Weighted PCA + UMAP	0.10	0.09	0.34	0.07	251.80	138.10
Kernel PCA	-0.01	0.01	0.02	0.05	374.60	276.10
Kernel PCA to UMAP	0.32	0.33	0.40	0.39	378.10	279.60

Table 1: Comparison of clustering algorithms and run time across dimension reduction methods for chromosome -filtered and -unfiltered data. Adjusted Rand Index (ARI) was a metric used to evaluate cluster performance on four clusters.

Conclusion

We analyzed chromatin contact data using hybrid models of PCA, UMAP, and weighted-UMAP applied on cell contact matrices and summary statistics to uncover structural patterns across cell types. As shown in Fig. 1, dimensionality reduction using PCA captured major sources of variance and enabled interpretable visualizations of chromosome-specific contributions to the data. Applying UMAP to the PCA-reduced data improved cluster separation, suggesting enhanced sensitivity to subtle structural differences and non-linear relationships, as expected with genomic data. Moreover, as shown in Fig. 5, the biplot captured nonlinear structures more effectively compared to PCA on the binarized contact matrix data alone. This suggests that the nonlinear mapping of contract matrices to summary-statistics allowed PCA to better represent the global structure of the data. Furthermore, applying UMAP to the KPCA-reduced data revealed clear separation of HAP1, HeLa, and K562 cell types, and significantly improved the ARI score. These results suggest the nonlinear structure of the data was present in the 50-dimensional KPCA-reduced space, UMAP yielded strong results in handling this non-linearity, improving both cluster compactness and separation between cell-types in 2D plots.

Visualizing the PC loadings as chromosome-block heatmaps shows that certain chromosomes contribute disproportionately to the variance, suggesting locus-specific changes in 3D genome structure. Interestingly, after filtering out chromosomes whose features on average contribute to less variance in the PC1 loading, the cluster performance remains relatively constant, while the run-time is approximately halved across all the evaluated dimensionality reduction techniques. This suggests that many chromosomes do not convey information critical to distinguishing between the four cell types studied and can be removed in favour of computational-speed improvements.

Despite visually observing improved separation in Fig 3. using weighted-UMAP, weighting gene loci by their distance from chromosome centers led to worse clustering performance. This suggests that such distance-based weighting may obscure biologically meaningful patterns, or the hypothesis that peripheral loci pairs might drive contact variability between different cell types may be incorrect. In contrast, unweighted data preserved more of the cell-type-specific interactions, yielding higher clustering ARI scores despite less distinct visual separation on the plot.

After performing PCA, clustering performance improved when using spectral clustering compared to K-means clustering, suggesting non-linear structural properties within Hi-C data. Interestingly, spectral clustering showed a peak in clustering score when the algorithm expects 4 groups, suggesting that the cell types have group differences detectable in an unsupervised learning setting.

Some limitations include unbalanced sample sizes of cells from each type (GM12878 and K562 are significantly under-represented in the data), hindering the ability to observe clear separation and clustering by cell type. Additionally, outliers observed in Fig. 1 and Fig. 5 were not fully addressed by binarization or chromosome filtering. However, the application of PCA and UMAP effectively clustered the data, with no clear outliers visible in the resulting projections.

Our findings demonstrate that low-dimensional embeddings of chromatin contact matrix data can reveal differences across cell types, offering a foundation for future investigation. Specifically, we can integrate group sparse principal component analysis (GSPCA) to enhance chromosome level pattern recognition and explore how closed (coiled) the chromatin are through the use of time-series cell-cycle and epigenomic datasets, if accessible. Alternative weighting schemes, such as logarithmic scaling, could also be explored to balance the contribution of short and long-range contacts.

References

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology, 24*(6), 417–441.

MacQueen, J. B. (1967). *Some methods for classification and analysis of multivariate observations*. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 281–297). University of California Press.

McInnes, L., Healy, J., & Melville, J. (2018). *UMAP: Uniform manifold approximation and projection for dimension reduction*.

Shi, J., & Malik, J. (2000). *Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8), 888–905. https://doi.org/10.1109/34.868688

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86), 2579–2605.