Open access • Posted Content • DOI:10.1101/2020.11.12.379941

THUNDER: A reference-free deconvolution method to infer cell type proportions from bulk Hi-C data — Source link

Bryce Rowland, Ruth Huh, Zoey Hou, Ming Hu ...+2 more authors

Institutions: University of North Carolina at Chapel Hill, Northwestern University, Cleveland Clinic Lerner Research Institute, University of California, San Francisco

Published on: 12 Nov 2020 - bioRxiv (Cold Spring Harbor Laboratory)

Topics: Population

THUNDER: A reference-free deconvolution method to infer cell type proportions from bulk Hi-C data

Authors

Bryce Rowland[1], Ruth Huh[1], Zoey Hou[2], Ming Hu[3], Yin Shen[4], Yun Li[5,1,6]

[1]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, [2]Department of Engineering Sciences and Applied Mathematics, Northwestern University, Evanston, IL, [3]Department of Quantitative Health Sciences, Lerner Research Institute, Cleveland Clinic, Cleveland, OH, [4]Department of Neurology, University of California San Francisco, San Francisco, CA, [5]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC, [6]Department of Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, NC

Abstract

Hi-C data provide population averaged estimates of three-dimensional chromatin contacts across cell types and states in bulk samples. To effectively leverage Hi-C data for biological insights, we need to control for the confounding factor of differential cell type proportions across heterogeneous bulk samples. We propose a novel unsupervised deconvolution method for inferring cell type composition from bulk Hi-C data, the Two-step Hi-c UNsupervised DEconvolution appRoach (*THUNDER*). We conducted extensive real data based simulations to test *THUNDER* constructed from published single-cell Hi-C (scHi-C) data. *THUNDER* more accurately estimates the underlying cell type proportions when compared to both supervised and unsupervised deconvolution methods including CIBERSORT, TOAST, and NMF. *THUNDER* will be a useful tool in adjusting for varying cell type composition in population samples, facilitating valid and more powerful downstream analysis such as differential chromatin organization studies. Additionally, *THUNDER* estimates cell-type-specific chromatin contact profiles for all cell types in bulk Hi-C mixtures. These estimated contact profiles provide a useful exploratory framework to investigate cell-type-specificity of the chromatin interactome while experimental data is still sparse.

Introduction

Statistical deconvolution methods have been applied extensively to studies of gene expression and DNA methylation[1–5] to infer cell type proportions and estimate cell-type-specific profiles. Deconvolution methods infer clusters from observed data which can correspond to either cell types or cell states, and here we refer to both as cell types for

brevity. In epigenome-wide association studies (EWAS) where the individual-level signal is a mixture of methylation profiles from different cell types, it has become standard practice to control for inferred cell type proportions when analyzing heterogeneous samples.[6] As we accumulate chromatin interaction information from heterogeneous samples using recently developed technologies such as Hi-C at an increasing rate, there will soon be sufficient individual level data to conduct similar 3D-chromatin-interactome wide association studies (3WAS) or chromatin interactome QTL (iQTL) studies[7]. Similar to DNA methylation and gene expression, there is growing evidence from single-cell Hi-C (scHi-C) data of important cell-to-cell variability in spatial chromatin interaction[8–10]. In order to effectively garner insights from associations between chromatin interactions and phenotypes of interest or to identify genetic determinants underlying variations in 3D-chromatin-interactome across biological samples, future 3WAS or iQTL analyses must control for the almost inevitable confounding factor of differential cell type proportions across heterogeneous bulk samples. If not accounted for, we risk inducing an increased false positive rate by Simpson's Paradox[6,11]. However, to the best of our knowledge, there is no statistical deconvolution method which is capable of leveraging both intrachromosomal and interchromosomal contacts for deconvolution across multiple bulk Hi-C samples simultaneously.

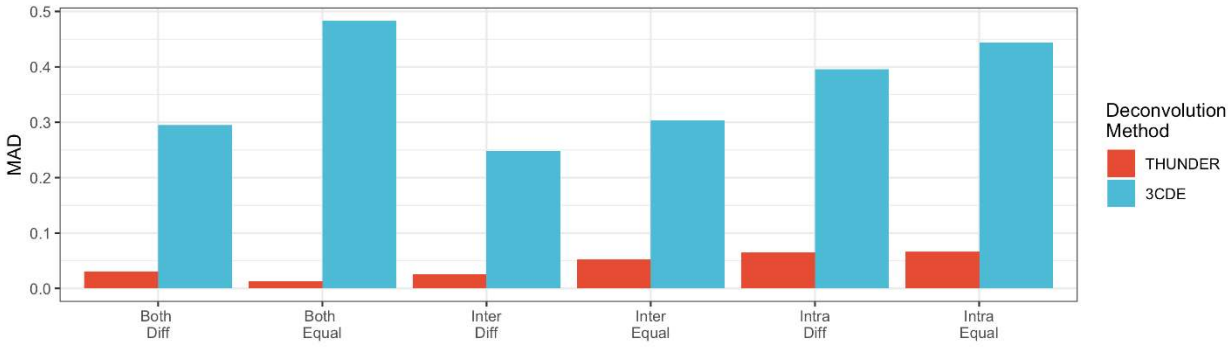
There exist two particular challenges of performing deconvolution in bulk Hi-C data: a lack of cell-type-specific Hi-C reference profiles and a two-dimensional (2D) nature to the aggregating unit of analysis, pairs of genomic loci (or bin-pairs). First, many deconvolution methods require cell-type-specific reference profiles for each cell type potentially present in a mixture, but these are widely unavailable for Hi-C data. Second, Hi-C data are usually

summarized in terms of pairs of genomic loci, but can be summarized at several different structural levels, such as A/B compartments, topologically associating domains (TAD's)[12], frequently interacting regions (FIRE's)[13,14], chromatin loops[15], interchromosomal contacts, and/or intrachromosomal contacts[16,17], and it is unclear which level(s) of measurement are most scientifically relevant. Whereas when deconvolving gene expression data, the aggregating unit of interest, the gene, is more obvious. Thus, a deconvolution method for Hi-C data must be robust to various representations of spatial data.

To address these challenges, we propose a non-negative matrix factorization (NMF) based Two-step Hi-c UNsupervised DEconvolution appRoach (*THUNDER*), to infer cell type proportions from bulk Hi-C data. *THUNDER* consists of a feature selection step and a deconvolution step, both of which rely on NMF. In the first step, we perform feature selection on the cell type profiles estimated from an initial deconvolution to identify informative bin-pairs in the mixture data (Figure 1a,b). We define informative bin-pairs as pairs of genomic loci with informative contact frequencies across cell types. In the second step, we perform deconvolution after subsetting the mixture matrix on informative bin-pairs (Figure 1c).

To the best of our knowledge, *THUNDER* is the first unsupervised deconvolution method for Hi-C data that integrates both intrachromosomal and interchromosomal contact information to estimate cell type proportions in multiple bulk Hi-C samples simultaneously. Two other matrix-based deconvolution approaches exist for Hi-C intrachromosomal contact matrices: 3CDE infers non-overlapping domains of chromatin activity in each cell type and uses a linear combination of binary interaction information at these domains to perform deconvolution.[18] Junier *et al.* put forth a method to infer

overlapping domains of chromatin activity as well as their mixture proportions.[19] Unlike *THUNDER*, neither method integrates information from interchromosomal bin-pairs. We tested 3CDE on our simulated bulk Hi-C mixtures, but found that it is almost impossible to apply in practice because it does not accommodate the inclusion of interchromosomal contacts and it requires across-sample cell type matching to align proportion estimates since it infers cell type proportions for each sample separately (Supplementary Figure 1).



Supplementary Figure 1. Performance of *THUNDER* and 3CDE on HAP1 and HeLa Simulated Mixtures. We see that in several simulations, 3CDE achieves near the maximum mean absolute deviation from true cell type proportions (0.5).We do not test 3CDE in further simulations because of its inability to handle multiple Hi-C samples simultaneously.

To the best of our knowledge, no software accompanies the work by Junier *et al.*[19] Carstens *et al.*[20] infers chromatin structure ensembles from bulk Hi-C contact information using a Bayesian approach but does not infer cell type proportions directly. In this work, we consider two other deconvolution methods developed for gene expression or methylation data. CIBERSORT is a standard reference-based deconvolution method developed for gene expression deconvolution.[2] TOAST is a recently published feature selection algorithm for gene expression or methylation data to select a pre-specified number of features while performing unsupervised deconvolution via NMF.[3]

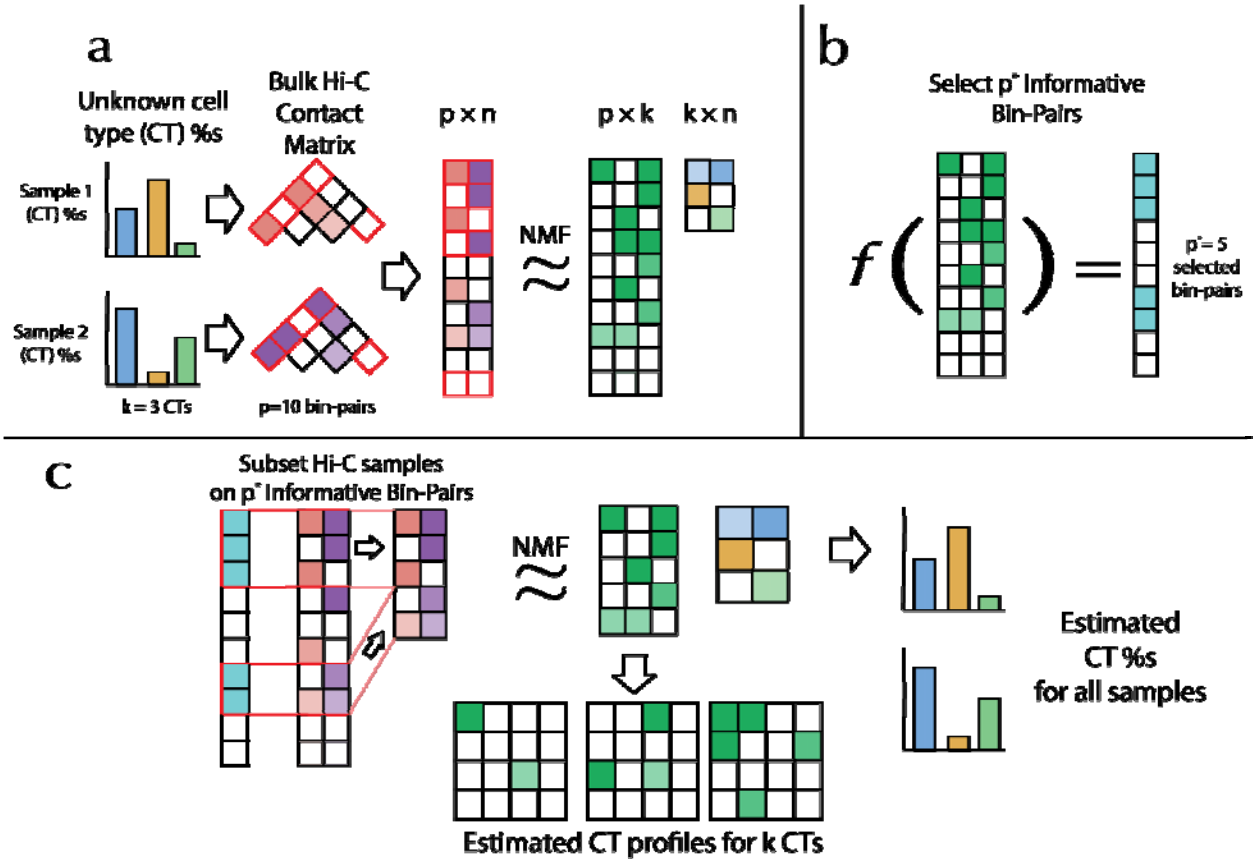


Figure 1. Overview of *THUNDER* Procedure. (a) Overview of nonnegative matrix factorization (NMF) in the context of bulk Hi-C data. Three underlying cell types each contribute to the observed contact frequencies in two bulk Hi-C samples. The first step of the *THUNDER* algorithm is to deconvolve our samples of Hi-C data into two estimated matrices: the cell type profile matrix and the proportion matrix. (b) *THUNDER* utilizes a feature selection algorithm specifically tailored to Hi-C data to analyze the contact frequency distribution of the bin-pairs in the cell type profile matrix in order to select informative bin-pairs (informative bin-pairs) for deconvolution. (c) In the final step of our algorithm, we subset the bin-pairs contained in our bulk Hi-C samples to only informative bin-pairs and perform NMF a second time. The proportion matrix is scaled to be estimates of the underlying cell type proportions in the bulk Hi-C samples. The cell type profile matrix estimates cell-type-specific contact distributions.

In order to test our proposed approach, we generate real data-based simulated bulk Hi-C datasets using scHi-C data from Ramani *et al.*[9] We generate bulk mixtures where the true cell type proportions in each sample are known (see **Methods**). We construct four such simulated Hi-C datasets: three mixtures of human cell types (HAP1 and HeLa mixtures; GM12878 and HeLa mixtures; GM12878, HAP1, and HeLa three-way mixtures),

and one mixture of mouse cell types (mouse embryonic fibroblast (MEF) and Patski fibroblast mixtures). In each of the experiments below, the input to *THUNDER* is a $p \times n$ matrix containing contact frequencies from bulk Hi-C mixtures. n is the number of mixture samples. p is the number of bin-pairs of Hi-C data with at least one non-zero entry across n samples. We performed normalization and downsampling of Hi-C data as described in **Methods**. Additionally, we assume that the underlying number of cell types, k, is known. We use simulated profiles based on scHi-C data derived from relatively homogeneous cell lines as additional input to the reference-based deconvolution methods as described in **Methods**. Finally, we run *THUNDER* five times for each experiment due to the random nature of the NMF algorithm and assess performance by averaging across the runs.

In all simulations, we compare *THUNDER*'s estimated cell type proportions to estimates from four other deconvolution methods: CIBERSORT, TOAST, NMF without feature selection (All Bin-pairs); and NMF with feature selection performed directly on the bulk Hi-C contact frequencies using the top 1,000 features by Fano factor.[2,3] When evaluating the performance of each method, we compute the mean absolute deviation (MAD) and Pearson correlation between the estimated and the true cell type proportions.

Results

THUNDER's Feature Selection Strategy

THUNDER's feature selection strategy was determined through a series of rigorous real-data based simulations across a variety of realistic scenarios in deconvolving bulk Hi-C samples containing mixtures of two human tumor cell types, HAP1 and HeLa. We test 11

published and novel NMF feature selection strategies (see **Supplementary Table 1** for definitions). In each simulation we consider the performance of the feature selection strategies across two Hi-C data read-outs, intrachromosomal contacts alone and interchromosomal contacts alone. Broadly, the tested feature selection strategies can be classified into two groups based on their input, either the bulk Hi-C contact frequencies or on the derived cell-type-specific profiles from an initial NMF deconvolution estimate. Strategies in the former group identify bin-pairs with high Fano Factor estimates across all samples. Strategies in the latter group identify informative bin-pairs with high cell-type-specificity and/or high variation across inferred cell types. Cell type specificity is measured by feature score within a bin-pair and across estimated cell types. Across-cell-type variation is measured by standard deviation within a bin-pair and across estimated cell types. For both metrics, we use empirical thresholds based on the distribution of these estimates across all bin-pairs for feature selection (see **Methods**). For fairness across the strategies based on the estimated cell type profiles, we use the same initial deconvolution estimate for all feature selection strategies.

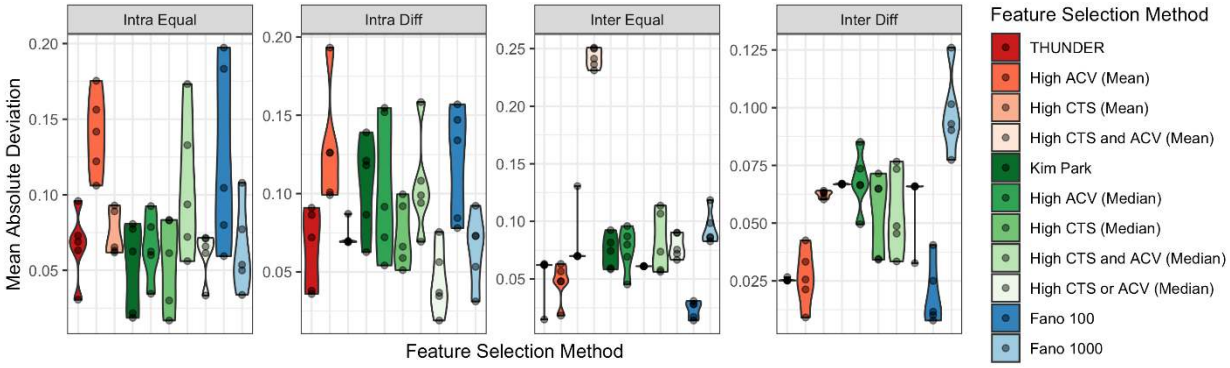
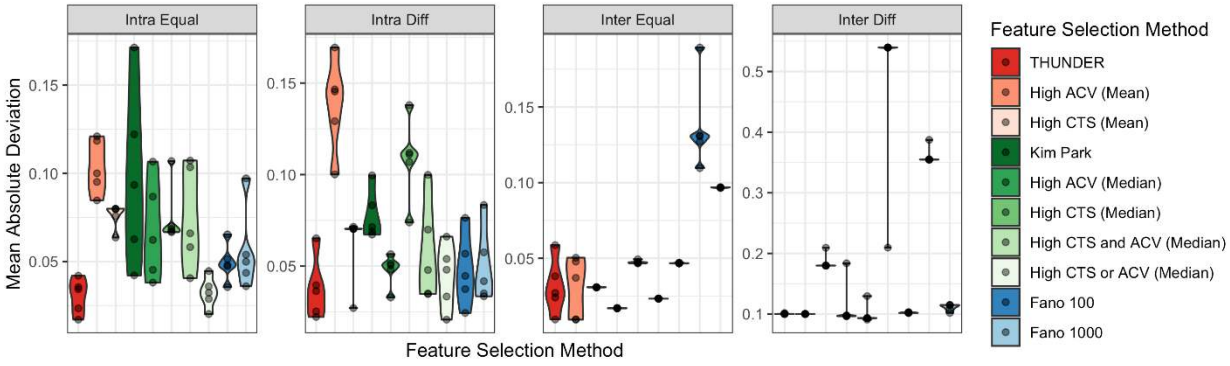


Figure 2. Performance of Feature Selection Strategies for Unsupervised Hi-C Deconvolution in HAP1 and HeLa Mixtures. We generate data using real-data based simulation of bulk Hi-C data with both intrachromosomal and interchromosomal contacts from scHi-C data from two human tumor cell lines HAP1 and HeLa. In "Intra" simulations, only intrachromosomal contacts are used, and in "inter" simulations only interchromosomal contacts are used. Additionally, in "equal" simulations the number of cells in each simulated bulk Hi-C sample are equal, and in "diff" simulations different bulk Hi-C samples have varying number of cells contributing to the mixture. We test 11 feature selection strategies including 8 novel feature selection strategies combining bin-pairs with high cell-type-specificity (CTS) and high across-cell-type variation (ACV). Colors are grouped such that the "reds" are feature score strategies analyzing the estimated cell-type-specific profiles using the mean across bin-pairs for thresholding, "greens" are feature score strategies analyzing the estimated cell-type-specific profiles using the median across bin-pairs for thresholding, and the "blues" are feature score strategies analyzing the input mixture matrix.



Supplementary Figure 2. Performance of Feature Selection Strategies for Unsupervised Hi-C Deconvolution in GM12878 and HeLa Mixtures.

The feature selection strategy that performs most robustly across all four simulations identifies bin-pairs with either high cell-type-specificity (CTS) or high across-cell-type variation (ACV) and adopts empirical thresholds based on the mean and standard deviation of measures across bin-pairs. We make this feature selection strategy the default and label it *THUNDER* in **Figure 2** and hereafter. While alternative feature selection strategies are less variable than *THUNDER* when deconvolving interchromosomal contacts only, they tend to perform poorly in other data contexts. Particularly sensitive are strategies which pre-specify the number of informative bin-pairs before deconvolution (Fano 100 and Fano 1000). We observe a similar sensitivity to *a priori* assumptions in feature selection strategies identifying bin-pairs with either high across-cell-type variation or high cell-type-specificity but not both. In practice, since the number of informative bin-pairs is unknown *a priori* and the preferred resolution of Hi-C data is unclear, using a non-robust feature selection strategy could lead to poor estimation of cell type proportions, and therefore incorrect inference when used in downstream analysis. *THUNDER*'s feature selection strategy uses a combination of both feature score and standard deviation to identify informative bin-pairs and is thus has more robust performance across input data types. *THUNDER*'s feature selection strategy performs well when the same strategies are tested on two cell mixtures of GM12878 and HeLa cells (**Supplementary Figure 2**).

Performance on HAP1 and HeLa Mixtures

We first evaluate the performance of *THUNDER* on real data-based simulated bulk Hi-C mixtures containing two human tumor cells lines, HAP1 and HeLa, by comparing estimated

and true cell type proportions across bulk samples (see **Methods**). In all simulation scenarios, *THUNDER* is either the best or close to the best across all methods compared. Notably, *THUNDER*, an unsupervised reference-free method, outperforms CIBERSORT, which uses cell-type-specific Hi-C reference profiles to perform supervised deconvolution. When deconvolving Hi-C data using both intrachromosomal and interchromosomal contacts where the number of underlying cells differs across samples ("Both Diff"), *THUNDER* reduces the MAD between estimated and true cell type proportions by 57% and 48% relative to CIBERSORT and NMF with no feature selection, respectively. In this "Both Diff" scenario, *THUNDER* performs comparably to the most recently proposed TOAST method. However, *THUNDER* clearly outperforms TOAST when only using intrachromosomal bin-pairs. *THUNDER* reduces MAD relative to TOAST by 44% and 35% when the number of cells in the mixture is equal and different across samples, respectively.

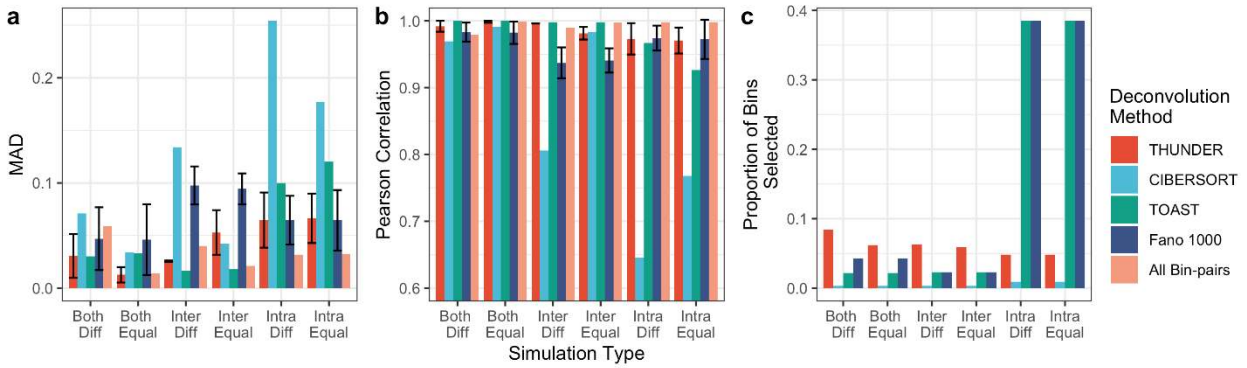
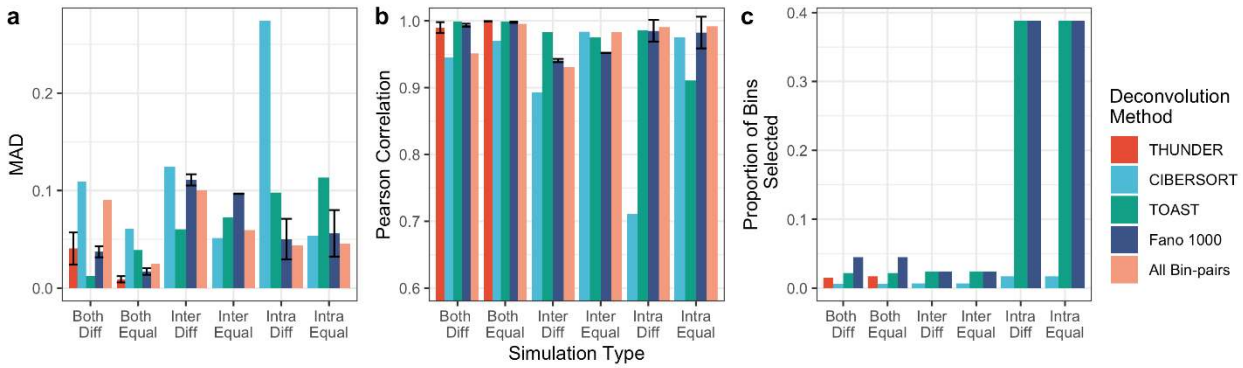


Figure 3 Performance of Deconvolution Methods on Mixtures with Two Human Tumor Cell Types HAP1 and HeLa. *THUNDER* and FANO 1000 are plotted with error bars (because NMF is a random algorithm. Simulation set up information detailed on x-axis including bin-pairs included in mixture ("Inter", "Intra" or "Both") and number of underlying cells across samples ("Equal" or "Diff"). (a,b) Average mean absolute deviation (MAD) and Pearson correlation comparing the true underlying cell

type proportions to the simulated true proportions across all simulations. Lower MAD and higher Pearson correlation indicates better performance. (c) Proportion of bin-pairs selected by deconvolution methods which perform feature selection.



Supplementary Figure 3 Performance of Deconvolution Methods on Mixtures with Two Human Cell Types GM12878 and HeLa.

In several simulations ("Both Equal", "Inter Equal", "Intra Diff", "Intra Equal"), NMF with no feature selection (All Bin-pairs) performs comparably to or even outperforms *THUNDER*. However, NMF with no feature selection performs poorly in more realistic scenarios with different number of cells across samples ("Inter Diff", "Both Diff"). In contrast, *THUNDER* is robust, performing either the best or close to the best across all settings.

When using only interchromosomal contacts, TOAST outperforms *THUNDER*, but *THUNDER* more efficiently integrates information from interchromosomal contacts with intrachromosomal signals when both contact types are deconvolved together. The human tumor cell lines, HAP1 and HeLa, demonstrate known interchromosomal translocations due to fusion events as the cancer cells mutate. Therefore, we expect all deconvolution methods to perform excellently in the interchromosomal only simulations due to strong differentiating signals between cell types. However, *THUNDER* deconvolves intrachromosomal and interchromosomal bin-pairs together more effectively than TOAST.

TOAST regresses in MAD by 89% and 78% when considering both intrachromosomal and interchromosomal contacts compared to only interchromosomal contacts where the number of cells across mixtures are equal and differ, respectively. However, *THUNDER* improves in MAD by 76% when using both interchromosomal and intrachromosomal contacts relative to only interchromosomal contacts when the number of cells across samples is equal, and regresses in performance by only 19% when the number of cells across samples differs.

To ensure that the deconvolution methods are not detecting artificial signals due to the near-haploid structure of the HAP1 cells in the bulk Hi-C mixture, we additionally test the same set of methods on a mixture of GM12878 and HeLa cells, noting that the number of cells in each sample is far fewer in these simulations due to the much smaller number of GM12878 cells with scHi-C data (see **Methods**). *THUNDER* again manifests the best performance (see **Supplementary Figure 3**).

Comparison of *THUNDER* and CIBERSORT on mice cell lines

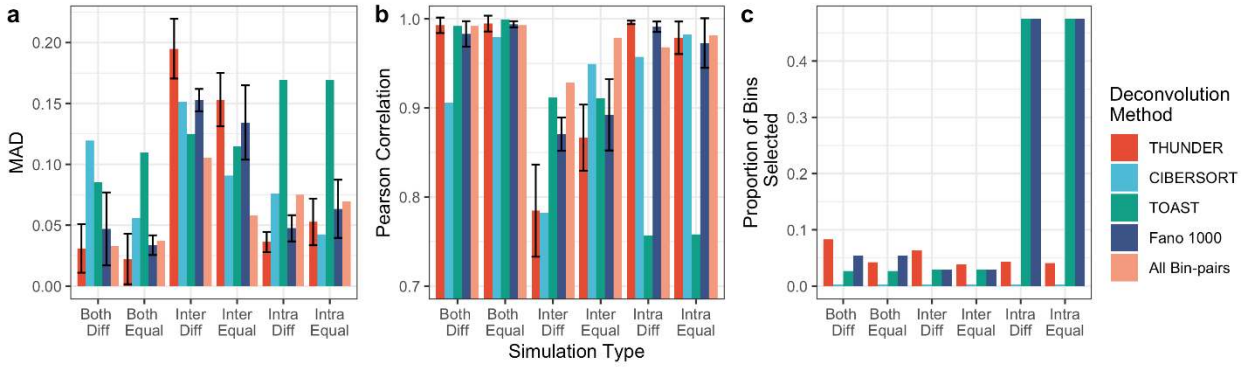


Figure 4 Performance of Deconvolution Methods on Mixtures with Two Mouse Cell Types MEF and Patski. THUNDER and FANO 1000 are plotted with error bars (because NMF is a random algorithm. (a,b) Average mean absolute deviation (MAD) and average Pearson correlation comparing the true underlying cell type proportions to the simulated true proportions across

simulations. Lower MAD and higher Pearson correlation indicates better performance. (c) Proportion of bin-pairs selected by deconvolution methods which perform feature selection.

To evaluate *THUNDER*'s performance when no tumor cells are involved and thus with no obvious informative cell-type-specific interchromosomal signatures, we simulate bulk Hi-C mixtures using real scHi-C data from two mouse embryonic cell lines, MEF and Patski. *THUNDER* outperforms all other methods tested in three of the four simulations using both types of bin-pairs or exclusively intrachromosomal bin-pairs. Additionally, we show that *THUNDER*'s estimated cell type proportions are closer to the truth when deconvolving intrachromosomal and interchromosomal contacts together rather than intrachromosomal contacts alone, even when performance using interchromosomal contacts alone is poor.

When considering interchromosomal and intrachromosomal bin-pairs together with differing number of cell types ("Both Diff"), *THUNDER* results in a 74% reduction in MAD compared to CIBERSORT, a 64% reduction compared to TOAST, and a 6% reduction compared to NMF with no feature selection. Similarly, *THUNDER* has the lowest MAD of all methods considered when deconvolving intrachromosomal contacts only with differing number of cell types ("Intra Diff" in **Figure 4**).

All methods incur a uniform increase in MAD when considering only interchromosomal bin-pairs (Figure 2a,b). We believe that increase in MAD (in contrast to the results from the HAP1 and HeLa simulations) is due to a lack of distinguishing cell type interchromosomal profiles in the MEF and Patski cell lines. However, we observe a general improvement in all deconvolution methods when analyzing interchromosomal and intrachromosomal bin-pairs together rather than exclusively using intrachromosomal contact information to perform deconvolution. In particular, using both data types together

when the number of underlying cells differ ("Both Diff") results in a 14% reduction in MAD for *THUNDER* compared to using only intrachromosomal bin-pairs. *THUNDER* effectively leverages both interchromosomal and intrachromosomal contacts to estimate underlying cell type proportions, improving on previous Hi-C deconvolution approaches which only consider intrachromosomal contacts.

THUNDER – Three Cell Types

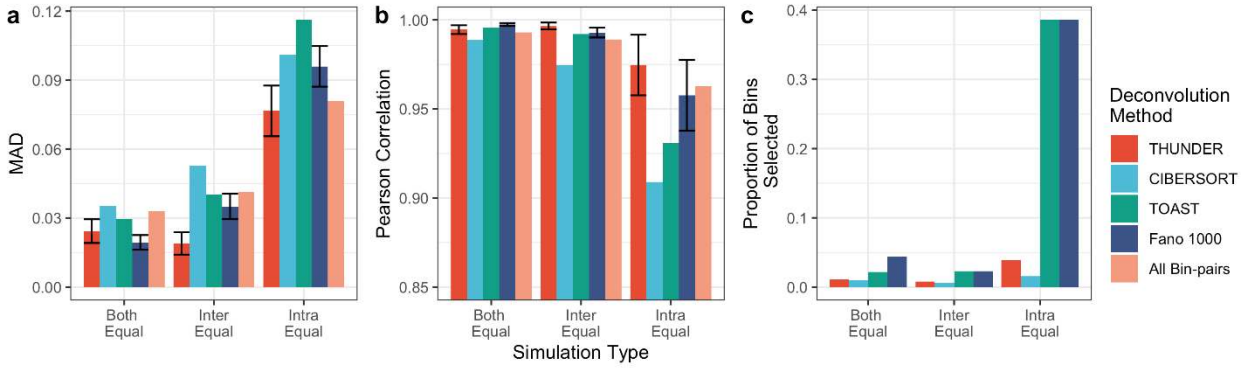


Figure 5 Performance of Deconvolution Methods on Mixtures with Three Human Cell Types: HAP1, HeLa, and GM12878. THUNDER and FANO 1000 are plotted with error bars ( ) because NMF is a random algorithm. (a, b) The average mean absolute deviation (MAD) and average Pearson correlation comparing the true underlying cell type proportions to the simulated true proportions across simulations. Lower MAD and higher Pearson correlation indicates better performance. (c) Proportion of bin-pairs selected by deconvolution methods which perform feature selection.

Finally, we evaluate the performance of the deconvolution methods in a more complicated scenario by generating samples from three human cell lines GM12878, HAP1, and HeLa. Here, we do not consider simulation settings where the number of cells varies across samples due to restrictions in the total number of cells after downsampling as described in **Methods**. *THUNDER* outperforms CIBERSORT, TOAST, and NMF with no feature selection

in all simulation settings considered. For example, when analyzing Hi-C mixtures with both intra- and interchromosomal contacts ("Both Equal"), *THUNDER* selecting ~1% of the total bin-pairs results in a 26% reduction in MAD from NMF with no feature selection, a 18% reduction in MAD from TOAST, and a 31% reduction in MAD from CIBERSORT (**Figure 5a**). As in the mouse cell line simulations, the addition of interchromosomal contacts improves performance uniformly across all deconvolution approaches as opposed to analyzing intrachromosomal contacts alone. *THUNDER* using intrachromosomal and interchromosomal bin-pairs together reduces MAD by 68% compared to using intrachromosomal contacts alone.

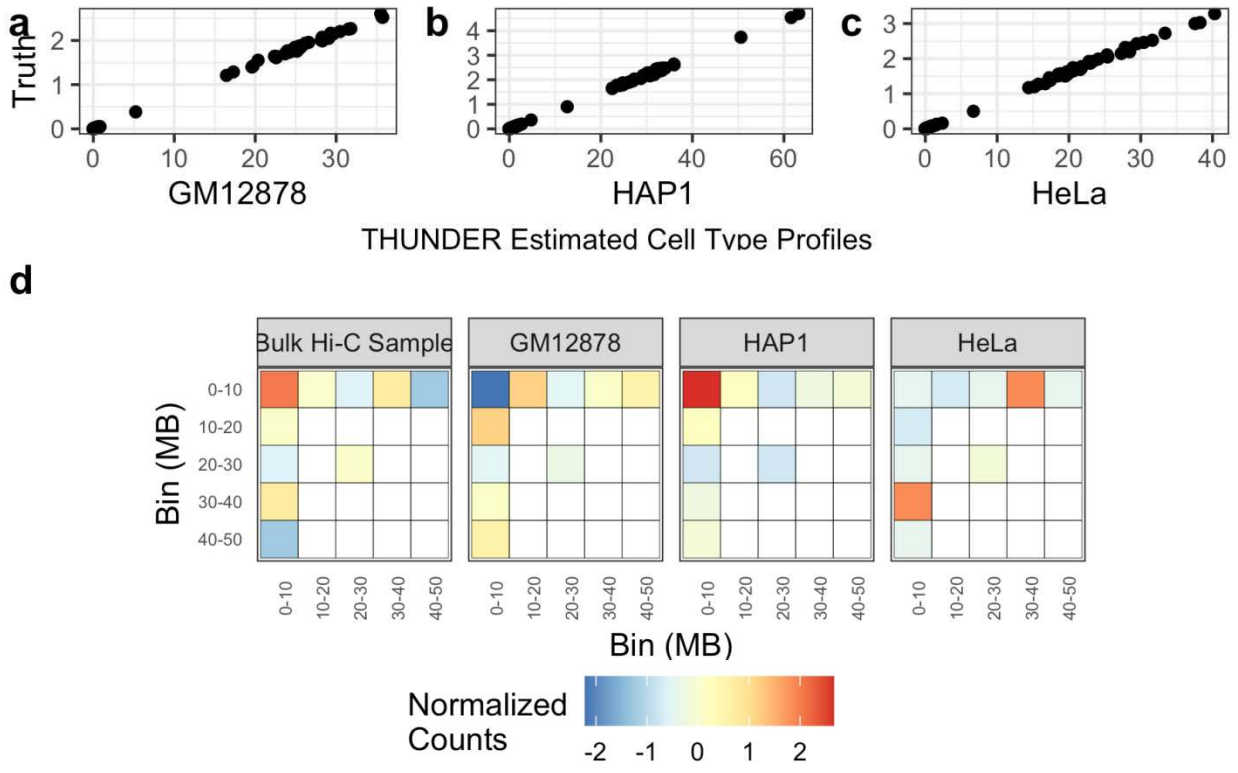


Figure 6 THUNDER Accurately Estimates Underlying Cell Type Proportions in Bulk Hi-C Data and Visualizes Cell Type Specific Hi-C Profiles. Results are taken from simulations using three human cell types: GM12878, HAP1, and HeLa. (a-c) Estimated meta-contact frequencies from the cell type profile matrix for Hi-C bin-pairs are highly correlated with contacts in pure simulated

samples. The cell type profile matrix is a good proxy for cell-type-specific Hi-C data. (d) Estimated cell type contact profiles at Intrachromosomal informative bin-pairs on chromosome 21 from real data based simulation of bulk Hi-C data. THUNDER cell type profile estimates identify regions of across-cell type variation which would have been missed if only bulk Hi-C data were analyzed.

In addition to providing accurate estimates of the underlying cell type proportions in bulk Hi-C data, *THUNDER* provides estimates of cell-type-specific interaction profiles for all cell types in the bulk Hi-C samples. Since Hi-C data is a population averaged estimate of contact frequency, cell-type-specific profiles estimated by *THUNDER* provide insights into differentiating patterns of interactions. Considering only *THUNDER* deconvolution estimates with the lowest MAD from the 3 human cell type simulations using both intra- and interchromosomal contacts, *THUNDER*'s cell-type-specific Hi-C interaction estimates are highly correlated with "true" cell-type-specific contact frequency where the "truth" is derived by aggregating information across scHi-C data from the same cell line (Figure 6a-c). When properly scaled, *THUNDER*'s cell-type-specific interaction estimates are a reasonable proxy for cell-type-specific chromatin activity at informative bin-pairs. Fig 6d displays cell-type-specific profiles estimated by *THUNDER* via heatmap of intrachromosomal contacts in a 50MB region on chromosome 21. Each cell type analyzed demonstrates distinct regions of both increased and decreased contact frequency compared to the other cells in the sample (Figure 5d) [21].

Computational Speed on Real Hi-C data.

We assess *THUNDER*'s computing performance on intrachromosomal Hi-C contact frequencies at 10Kb resolution from five individuals from Yoruba in Ibadan, Nigeria (YRI).[7] Note that 10Kb resolution results in 38,343,298 unique intrachromosomal bin-pairs to consider, ranging 380,000 to 3.5 million bin-pairs per chromosome. To obtain cell type proportion estimates genome-wide using *THUNDER*, we first perform feature selection by chromosome, then concatenate the selected features across chromosomes as input for the final deconvolution estimate. *THUNDER*'s mean computing time is 3.4 hours and has a mean memory usage of 57GB per chromosome. The final genome-wide estimation step to obtain cell type proportions took 2.5 hours and 18GB of memory (Supplementary Table 2). For the 5 YRI samples, 693,771 (~2%) bin-pairs are selected as informative bin pairs. Similar summaries are presented for deconvolution estimates with 3 and 10 YRI samples respectively (Supplementary Table 1,3). One advantage of *THUNDER*'s feature selection method when analyzing genome-wide Hi-C data is the ease with which it can be parallelized by subsetting the original input mixture matrix in smaller regions than by chromosome, then concatenating Hi-C data for the final cell type proportion estimation step. This run time and memory usage serves as an upper limit on the computational costs of running *THUNDER*, as 10Kb is one of the finest resolutions of Hi-C data currently analyzed in practice. We expect *THUNDER*'s computing cost to scale approximately linearly both with the number of bin pairs analyzed and the number of samples included (Supplementary Tables 1-3).

Chr	Duration (h)	Memory (GB)	Bin-Pairs	Selected Bin-Pairs
THUNDER Step 1				
chr1	4.7	85.1	2,898,116	57,611
chr2	3.7	96.7	3,211,342	63,152
chr3	4.5	70.3	2,742,971	54,380
chr4	4.4	73.1	2,653,508	52,331
chr5	4.0	73.1	2,457,251	48,239
chr6	3.8	54.6	2,334,790	46,153
chr7	2.5	52.1	1,972,510	39,288
chr8	2.3	51.9	1,926,609	38,021
chr9	1.5	44.0	1,443,866	28,720
chr10	3.2	48.5	1,681,822	33,231
chr11	3.7	51.5	1,727,598	34,530
chr12	3.6	51.7	1,752,414	35,044
chr13	2.5	37.6	1,364,537	26,453
chr14	1.3	41.2	1,170,924	23,367
chr15	1.0	31.0	996,396	20,132
chr16	1.5	29.3	832,338	16,141
chr17	1.8	29.6	841,985	17,422
chr18	0.9	32.2	1,025,512	20,399
chr19	0.9	22.1	468,526	9,114
chr20	0.9	26.4	730,999	14,688
chr21	0.7	22.8	441,879	8,468
chr22	0.3	19.5	339,585	6,887
THUNDER Step 2				
Genome-Wide	0.7	16.2	693,771	NA

Supplementary Table 1. Computational Performance on 3 YRI Samples of 10Kb Resolution Hi-C Data.

Chr	Duration (h)	Memory (GB)	Bin-Pairs	Selected Bin-Pairs
THUNDER Step 1				
chr1	5.4	97.3	3,174,333	65,265
chr2	7.2	103.2	3,507,051	68,506
chr3	6.3	92.4	2,988,427	61,700
chr4	5.1	89.1	2,893,327	59,152
chr5	4.0	85.7	2,677,352	53,739
chr6	6.9	81.2	2,545,972	50,167
chr7	6.1	69.6	2,163,857	44,196
chr8	3.3	59.1	2,105,779	41,971
chr9	2.5	49.0	1,580,008	32,001
chr10	3.2	59.4	1,848,124	36,801
chr11	3.1	57.5	1,896,637	37,615
chr12	2.9	62.2	1,931,846	38,602
chr13	2.3	49.2	1,483,383	29,353
chr14	1.9	45.3	1,278,870	25,480
chr15	2.1	35.7	1,094,740	22,468
chr16	2.4	34.5	924,458	17,667
chr17	3.8	37.3	935,763	19,203
chr18	1.5	42.7	1,120,797	21,741
chr19	1.5	23.5	524,543	9,540
chr20	1.3	30.0	802,176	15,743
chr21	1.1	25.2	485,244	9,076
chr22	0.6	21.1	380,611	7,713
THUNDER Step 2				
Genome-Wide	2.5	18.2	767,700	NA

Supplementary Table 2. Computational Performance on 5 YRI Samples of 10Kb Resolution Hi-C Data.

Chr	Duration (h)	Memory (GB)	Bin-Pairs	Selected Bin-Pairs
THUNDER Step 1				
chr1	12.6	110.5	3,538,635	68,709
chr2	15.4	160.2	3,903,593	78,890
chr3	20.3	103.8	3,306,003	68,256
chr4	17.4	106.6	3,201,164	65,000
chr5	11.0	114.4	2,967,778	58,602
chr6	9.8	86.9	2,818,330	56,838
chr7	12.9	93.5	2,426,518	50,105
chr8	11.8	89.6	2,343,805	46,794
chr9	6.6	73.7	1,761,728	34,939
chr10	11.4	80.4	2,070,850	41,651
chr11	9.5	85.1	2,117,742	42,870
chr12	7.6	84.6	2,129,312	41,521
chr13	12.1	63.5	1,638,718	31,997
chr14	5.2	62.1	1,431,155	28,821
chr15	6.3	53.3	1,234,104	24,959
chr16	5.1	47.6	1,060,098	20,798
chr17	3.7	43.9	1,075,081	21,785
chr18	4.8	54.0	1,245,372	25,415
chr19	3.6	27.9	616,312	10,771
chr20	4.7	39.9	900,160	18,390
chr21	3.8	28.1	545,011	10,082
chr22	1.5	27.1	440,207	8,700
THUNDER Step 2				
Genome-Wide	4.1	28.6	855,893	NA

Supplementary Table 3. Computational Performance on 10 YRI Samples of 10Kb Resolution Hi-C Data.

Discussion

THUNDER is the first unsupervised deconvolution method for Hi-C data to integrate both intrachromosomal and interchromosomal contact information to estimate cell type proportions in multiple bulk Hi-C samples simultaneously. Across all simulations, *THUNDER*'s accuracy in estimating cell type proportions either exceeds or nearly matches all other Hi-C deconvolution approaches, including the supervised method CIBERSORT and

the recently proposed TOAST. We find that even in non-cancerous cell lines, the inclusion of sparse interchromosomal contact information in addition to intrachromosomal contacts improves deconvolution performance. *THUNDER* also provides an approach to infer cell-type-specific contact frequency from bulk Hi-C data. Importantly, *THUNDER*'s feature selection strategy for identifying informative bin-pairs before deconvolution improves performance relative to NMF with no feature selection.

We have shown that *THUNDER* successfully estimates underlying cell type proportions when deconvolving intrachromosomal and interchromosomal contacts together. In most cell types, we have more reliable Hi-C data at a much larger number of intrachromosomal bin-pairs compared to interchromosomal bin-pairs. For this reason, previous methods to deconvolve Hi-C data restricted their estimation to these intrachromosomal contacts. However, even in the simulations with no strong interchromosomal signatures, *THUNDER*'s performance improves when integrating interchromosomal and intrachromosomal data for deconvolution relative to only using intrachromosomal contacts, suggesting value in including interchromosomal contacts in bulk Hi-C data analysis despite questions regarding the robustness in quantifying interchromosomal contacts.

Additionally, we demonstrate that *THUNDER* successfully estimates cell-type-specific interaction profiles, and these profiles can be used to further examine variability across cell types in chromatin organization while there remains a paucity of cell-type-specific Hi-C data for cell types present in primary tissues (**Figure 6**). *THUNDER*'s feature selection strategy identifies a parsimonious subset of the overall number of bin-pairs which most differentiate the cell types. Thus, the estimated cell type profile matrix serves a dual

purpose: identifying informative bin-pairs from the large input feature space (dimension reduction) and accurately estimating relative contact frequency at informative bin-pairs.

One possible application of these cell-type-specific contact profiles could be in fine-mapping of GWAS variants in non-coding regions of the genome. Genome-wide association studies (GWAS) have identified over 200,000 unique associations between single-nucleotide polymorphisms (SNPs) and common diseases or traits of interest[22]. However, the majority of these SNPs reside in regulatory regions where little is understood about their underlying functional mechanisms which has limited the adoption of variant-trait associations into revealing molecular mechanisms and further into transforming clinical practice. Functional annotation of GWAS results are often most relevant at a cell-type-specific level due to important variability across cell types[23]. By further understanding the cell-type-specific interactome via *THUNDER*'s estimated profiles, we can more appropriately link putatively causal variants identified by GWAS to the target genes on which they act.

While we have presented real-data based simulation results for Hi-C data here, the *THUNDER* algorithm could easily be modified to other data types which capture the spatial organization of chromatin such as HiChIP/PLAC-seq data (HP data), which couple standard Hi-C with chromatin immunoprecipitation to profile chromatin interactions anchored at genomic regions bound by specific proteins or histone modifications, with reduced cost and enhanced resolution[24,25]. Used in concert with methods to identify long-range chromatin interactions from HP data[26], our method is anticipated to efficiently leverage interchromosomal contacts jointly with high quality intrachromosomal contacts to estimate underlying cell type proportions. The robustness of our feature selection strategy

and subsequent deconvolution performance warrant future interrogation in the setting of HP data.

There are two primary limitations of our simulation approach both attributable to the current quality of scHi-C data. First, due to the number of cells present in current scHi-C datasets and the library size, our analysis was limited to a coarse resolution of 10Mb bins when generating our synthetic bulk Hi-C data. However, we find that *THUNDER* still performs exceedingly well in estimating true cell type proportions even at coarse resolution. Secondly, the number of cell types and the overall coverage of the genome with our synthetic bulk Hi-C data are both much lower than one would expect in a realistic sample of bulk Hi-C data. As more scHi-C data becomes available, we hope to continue to test *THUNDER* in different real-data based scenarios which may be more realistic in terms of Hi-C data's read-depth.

Conclusion

To summarize, we present *THUNDER*, an unsupervised deconvolution approach tailored to the unique challenges of deconvolving Hi-C data. *THUNDER* accurately estimates cell type proportions in bulk Hi-C data regardless of the readout of Hi-C data considered. *THUNDER*'s biologically motivated feature selection approach performs well when considering both mouse and human cell lines. We have demonstrated the computational efficiency of the method through our analysis of 10Kb resolution Hi-C data. Finally, the estimated cell-type-specific chromatin interactome profiles are valuable for identifying bin-pairs which interact differentially across cell types.

One extension of *THUNDER* would be to include an additional step where the

number of cell types in the mixture is estimated, but we do not consider the approach here. In practice it would also be challenging to associate the deconvolved proportions back to specific cell types without careful experimental constructions of the proportions in each sample. This is due to the limited number of relevant cell-type-specific Hi-C profiles. Here, because we know the samples which are cell-type-specific in each simulation, we can map back the estimated cell type proportions and thus the cell-type-specific profiles to specific cell types in our simulation. This challenge becomes much more difficult if the number of cell types in a mixture is large and pure samples are not available, but is a common challenge for reference-free deconvolution approaches.

Accurately estimating underlying cell type proportions via *THUNDER* should be the first step in any individual-level differential analysis of bulk Hi-C data to control for the inevitable confounding factor of underlying cell type proportions. Additionally, *THUNDER* provides a unique tool to identify differentially interacting bin-pairs at the cell-type-specific level which can be associated with disease or phenotypes of interest. An R package for running THUNDER can be downloaded from https://github.com/brycerowland/thundeR.git. We anticipate *THUNDER* to become a convenient and essential tool in future multi-sample Hi-C data analysis.

Methods

THUNDER. In order to estimate the underlying cell type proportions found in bulk Hi-C datasets, we propose a Two Step Hi-C UNsupervised DEconvolution appRoach (*THUNDER*). As an unsupervised deconvolution method, *THUNDER* does not rely on any cell-type-specific reference profiles, making it particularly valuable in the current literature because

of the current paucity of cell-type-specific reference. In addition to estimating cell type proportions, *THUNDER* also infers cell-type-specific Hi-C interaction profiles which most differentiate the cell types in the mixture.

THUNDER consists of a feature selection step and a deconvolution step, both of which rely on non-negative matrix factorization (NMF, see below). In step one of *THUNDER*, we perform an initial NMF deconvolution estimate and perform feature selection using the decomposition to identify informative bin-pairs across cell types. In our deconvolution step, we reduce the original bulk Hi-C mixture matrix to only contain the informative bin-pairs identified by feature selection and then perform deconvolution a second time to estimate cell type proportions in each mixture and to infer cell-type-specific interactome profiles.

Consider an $p \; x \; n$ mixture matrix, V, where p is the number of bin-pairs and n is the number of samples. Let $p = p_{inter} + p_{intra}$ where p_{inter} and p_{intra} are the number of interchromosomal and intrachromosomal contacts respectively. We wish to deconvolve V in order to estimate the proportions of the k cell types the n samples, where k is known. The two steps of *THUNDER* are outlined below in the setting where V contains both interchromosomal and intrachromosomal contacts, but *THUNDER* can easily be applied to situations where only one data type is present.

First, we perform NMF on the $p \; x \; n$ matrix V to obtain the deconvolution estimate $V \approx W_1 H_1$ where W_1 is a $p \; x \; k$ matrix and H_1 is a $k \; x \; n$ matrix. W_1, the cell-type-specific profile matrix, contains estimates of cell-type-specific interaction profiles. H_1, known as the proportion matrix, is a $k \; x \; n$ matrix containing the estimates of cell type proportions when scaled appropriately. We compute summary statistics across all samples for each bin-pair

present in the rows of W_1 and then compare these statistics within the contact-types present in the mixture matrix to select bin-pairs which differ across cell types (detailed in ***THUNDER* Feature Selection** section below). For example, if we consider V to contain both intra- and interchromosomal contacts, we perform our feature selection strategy separately on the p_{intra} intrachromosomal contacts and on the p_{intra} interchromosomal contacts. After identifying p^* informative bin-pairs, we subset V on all informative bin-pairs regardless of contact type to form the reduced $p^* x\ n$ mixture matrix V^*. We then perform NMF on V^* to arrive at our final estimates , W^* (of dimension $p^*\ x\ k$) and H^* (of dimension $k\ x\ n$). Finally, we adjust the columns of H^* to sum to one to represent cell type proportions.

***THUNDER* Feature Selection.** The first step of *THUNDER* is to select informative bin-pairs which differentiate the cell types in the mixture. As described in the method outline above, we first deconvolve the input bulk Hi-C data matrix, V, into the estimates $V \approx W_1 H_1$. The rows of the cell type profile matrix, W_1, represent bin-pairs and the columns are inferred cell types present in the Hi-C mixtures. Feature selection is performed by computing summary statistics across cell types (columns of W_1) and within each bin-pair (rows of W_1), and then using empirically derived cutoffs to select informative bin-pairs in an unsupervised manner. Our feature selection strategy relies on two summary statistics, standard deviation and feature score. Both statistics are specific to each bin-pair and calculated across inferred cell types. Standard deviation allows selection of bin-pairs with high across-cell-type variation and feature score prioritizes bin-pairs with high cell-type-specificity.

Let $W_1(i,j)$ denote the element in the i^{th} row and j^{th} column of the cell-type-specific profile matrix W_1. Then $i = 1, \dots, p$ indicates bin-pair i. Let S_{intra} denote the set of all intrachromosomal bin-pairs. The derivation below is for intrachromosomal bin-pairs, but the feature selection algorithm is the same for interchromosomal variants. Standard deviation across cell types for base-pair i is defined as,

$$SD_i = \frac{1}{k-1}\sum_{j=1}^{k}\left(W(i,j) - \frac{1}{k}W(i,\cdot)\right)^2$$

Larger values of the standard deviation across cell types for a given bin-pair indicate greater variation in the estimated cell-type-specific chromatin activity.

Feature score across cell types for base-pair i is defined as follows.

$$(Feature\ Score)_i = 1 + 1/log_2(k)\sum_{j=1}^{k} p(i,j)log_2\big(p(i,j)\big)$$

where $p(i,\Omega)$ is the probability that the i-th pairwise bin contributes to cell type Ω , i.e. $p(i,\Omega) = \frac{W_1(i,\Omega)}{\sum_{j=1}^{k} W_1(i,j)}$. Feature scores range from [0,1] with higher scores representing bin-pairs with higher cell-type-specificity.

Consider,

$$\hat{\mu}_{SD,intra} = \frac{1}{|S_{intra}|}\sum_{\{i:\ i\in S_{intra}\}} SD_i$$

$$\hat{\mu}_{FS,intra} = \frac{1}{|S_{intra}|} \sum_{\{i:\, i \in S_{intra}\}} FS_i$$

$$\hat{\mu}_{SD,intra} = \frac{1}{|S_{intra}| - 1} \sum_{\{i:\, i \in S_{intra}\}} \left(SD_i - \hat{\mu}_{sd,intra}\right)^2$$

$$\hat{\mu}_{FS,intra} = \frac{1}{|S_{intra}| - 1} \sum_{\{i:\, i \in S_{intra}\}} \left(FS_i - \hat{\mu}_{sd,intra}\right)^2$$

Bin-pair i is defined to be an informative bin-pair if $FS_i > \hat{\mu}_{fs} + 3\hat{\sigma}_{fs}$ or $SD_i > \hat{\mu}_{sd} + 3\hat{\sigma}_{sd}$.

Interpretation of *THUNDER* Estimates. If we scale the elements of H^* so that the columns of the matrix sum to 1, we can interpret the elements as the estimates of the cell type proportions in the p mixture samples. This scaled matrix is only for estimation of the cell type proportions, and the unscaled H^* should be used in conjunction with W^* for estimating contact frequency.

We can interpret the columns of W^* as parsimonious cell-type-specific contact profiles. These parsimonious contact profiles identify the most differentially interacting regions of chromatin activity across the cells in the mixture. Due to the paucity of reference cell-type-specific Hi-C profiles because of challenges with sorting homogenous cells from primary tissue[27] or with scHi-C experiments[28], the estimated cell-type-specific contact profiles from *THUNDER* provide useful approximation to chromatin interactome profiles specific to major cell types present in bulk Hi-C samples.

Nonnegative Matrix Factorization. NMF has been used in many computational biology applications to cluster genes, discover cancer types using microarray data, and study functional relationships of genes[29–31]. For Hi-C data, V denotes the $p \, x \, n$ mixture matrix of bulk Hi-C samples with p bin-pairs and n columns of mixture samples. We let $k > 0$ be an integer specified for the rank of our W matrix. NMF seeks to find an approximation $V \approx WH$, where W and H are $p \times k$ and $k \times n$ non-negative matrices. In the context of deconvolving Hi-C data, k is the number of distinct cell types in the mixture sample and often $k \ll min(n, p)$. In our analysis, k is chosen *a priori*. We refer to W and H as the cell type profile and proportion matrices, respectively. Since *THUNDER* extends NMF by performing feature selection, we refer to the results of deconvolution via *THUNDER* as W^* and H^*.

The NMF problem can be solved by finding a local minimum for a given objective function under the constraint that the resulting deconvolution has non-negative matrices. We use the NMF R package[32] with the updates provided by Lee and Seung[33] to run 200 iterations with random initialization of the W and H matrices. Our estimated cell type profile and proportion matrices are from the factorization that outputs the smallest deviance according to the cost function. In this paper, we explore two different objective functions for *THUNDER*, the Frobenius norm and the Kullback-Leibler (KL) divergence (**Supplementary Results**). The two objective functions are defined as $\| V - WH \|^2 = \sum_{ij} \left(V_{ij} - (WH)_{ij}\right)^2$ and $D(V \| WH) = \sum_{ij} (V_{ij} log \left(\frac{V_{ij}}{(WH)_{ij}}\right) - V_{ij} + (WH)_{ij})$, respectively. The two formulations of NMF are solved by minimizing these cost functions with the constraint that the resulting matrices in the decomposition are non-negative, $W, H \geq 0$.

Note that these functions are not convex in both W and H, therefore we cannot find a global minimum, but it is possible to find a local minimum.

CIBERSORT. CIBERSORT is a supervised deconvolution method which is widely used in the gene expression literature to estimate cell type proportions in tissue samples based on reference gene expression profiles.[2] CIBERSORT deconvolves samples using an application of linear support vector machines. In all applications, we run CIBERSORT using their browser-based implementation of the software[34]. Using default parameters, we run each deconvolution for 500 iterations. We define our own "signature gene" files using the simulated pure samples described below noting that "signature gene" files are actually bin-pairs for Hi-C data, but borrowing from CIBERSORT terminology. The informative bin-pairs reported in each analysis are those created by the CIBERSORT portal.

TOAST. TOAST is a recently proposed unsupervised deconvolution and feature selection algorithm which iteratively searches for cell type-specific features and performs composition estimation.[3] We use the TOAST Bioconductor package version 1.0.0 using the default 1,000 features for deconvolution. Additionally, we use NMF with KL divergence function as the deconvolution engine of TOAST.

3CDE. 3CDE is a matrix-based deconvolution approach for bulk Hi-C data which infers non-overlapping domains of chromatin activity in each cell type from data and uses a linear combination of binary interaction information at these domains to deconvolve the contact frequency matrix.[18] We downloaded software from their Github page

(https://github.com/emresefer/3cde), and ran *3cdefrac.py* with default settings. We found that the results were not usable when deconvolving multiple samples with the same underlying cell types without additional feature matching algorithms (see **Supplementary Figure 1**).

Simulating Bulk Hi-C Data. In order to simulate realistic bulk Hi-C mixtures on which to test *THUNDER* and alternative methods, we mix scHi-C data from Ramani et al.[9] The cellular indices data from the Ramani paper were downloaded from GSE84920 which included 6 libraries: ML1, ML2, ML3, ML4, PL1 and PL2. For our analysis, we use data from all libraries except ML4. These libraries are comprised of scHi-C data from five distinct human and mouse cell lines. We observe differences in the scHi-C read depth across the five libraries, and thus adopt a downsampling approach to account for the difference. Within each cell, we follow the same preprocessing procedure as outlined in Ramani *et al.*[9] Specifically, cellular indices with fewer than 1000 unique reads, a *cis*:*trans* ratio less than 1, and cells with less than 95% of reads aligning uniquely to either the mouse or human genomes are filtered out before analysis. Additionally, we remove reads whose genomic distance was <15Kb due to self-ligation, and only considered unique reads. For the four libraries with HAP1 and HeLa cells (ML1, ML2, PL1 and PL2), cellular indices were discarded where the proportion of sites where the non-reference allele was found was between 57% and 99%.

To account for varying levels of single-cell sequencing depth across cell lines and libraries, we consider only cells with filtered reads greater than the 20^{th} quantile and less than the 90^{th} quantile of reads and across all libraries and cell types considered in the

simulated mixture sample. We then downsample each cell via multinomial sampling to the number of contacts in the cell with the fewest number of contacts across all cell types considered in the sample. With the filtered and downsampled scHi-C data, we construct contact matrices at three levels of data representation at 10Mb bin-pair resolution: interchromosomal contacts only, intrachromosomal contacts only, and both interchromosomal contacts and intrachromosomal contacts together. The total number of cells in each mixture sample is equal to the smallest number of cells present after the filtering step across cells in the mixture sample.

In large part, the 10Mb window choice was limited by the library size of current scHi-C datasets and sparsity of contacts from which to generate synthetic bulk Hi-C datasets such that the true cell type proportions are known. Additionally, we report from our computation test on 10Kb resolution Hi-C data that *THUNDER* scales up to the much larger feature space of finer resolution Hi-C data. As single-cell technologies improve, we will be able to test Hi-C deconvolution methods at finer data resolutions where truth is known.

HAP1 and HeLa mixtures. Using the four human scHi-C libraries (ML1, ML2, PL1, and PL2) from Ramani *et al.*, we mix HAP1 and HeLa single-cell contacts to make our bulk Hi-C datasets at each of the three levels of data representation: interchromosomal contacts only, intrachromosomal contacts only, and both interchromosomal and intrachromosomal contacts.[9] For each representation of Hi-C data, we generate bulk Hi-C dataset under three scenarios: pure bulk samples, bulk samples which are a mixture of the two cell types with an equal number of total cells in each sample, and bulk samples which are a mixture of the

two cell types in which the number of cells in each sample varies. First, we generate 6 pure bulk samples, 3 samples which are a bulk collection of 855 HELA cells and 3 samples which are a bulk collection of 855 HAP1 cells. Second, we generate 11 bulk samples which are mixtures of HeLa and HAP1 cells at the sequence of (0% - 100% by an increment of 10%,) HAP1 cells, each with a total number of 855 cells. Finally, we generate 11 bulk samples, 6 of which are mixtures of HeLa and HAP1 cells at the sequence of (0% - 100% by an increment of 10%) HAP1 cells with a total number of 855 cells, and 5 of which are mixtures of HeLa and HAP1 cells at the sequence of (10-90% by an increment of 10%) HAP1 cells with a total number of 428 cells. In all cases, the resulting mixture matrices have rows representing pairwise contact bin-pairs and columns representing the bulk samples. These bulk samples are then normalized by adjusting for the total number of contacts which pass the filtering criteria outlined above, and multiplied by a million. The final mixture matrix represents the reads per million that map to each pairwise contact bins.

GM12878 and HeLa mixtures. The simulation approach is the same as above for GM12878 mixtures with the key difference being the number of cells which pass QC for each cell type. Here, we downsample the two cell lines to 407 cells. The number of samples and sample proportions in the GM12878 and HeLa mixtures are the same as above.

GM12878, HAP1, and HeLa mixtures. To test how *THUNDER* performs with more than two cell types in a mixture, we extend the above simulation to consider mixtures of HAP1, HeLa, and GM12878 cell lines. Since only 407 GM12878 cells pass our filtering procedure, we do not generate bulk mixture samples under the scenario where the cell types are different.

First, we generate 9 pure bulk samples which are mixtures of each cell type consisting of 407 cells. Second, we generate 12 bulk samples (3 pure samples and 9 mixture samples from comprised of 2 different mixture permutations) which are mixtures of the 3 cell lines at the proportions given in **Supplementary Table 4**. These proportions are a subset of those used by Shen-Orr and Tibsherani in their simulated mixture data.[1]

Table 1: Mixing proportions for GM12878, HAP1, and HeLa mixtures.

Sample Number	GM12878	HAP1	HeLa
1	1.00	0.00	0.00
2	0.00	1.00	0.00
3	0.00	0.00	1.00
4	0.70	0.25	0.05
5	0.70	0.05	0.25
6	0.05	0.70	0.25
7	0.05	0.25	0.70
8	0.25	0.05	0.70
9	0.25	0.70	0.05
10	0.45	0.45	0.10
11	0.45	0.10	0.45
12	0.10	0.45	0.45

Supplementary Table 4. Mixing Proportions for GM12878, HAP1, and HeLa Simulations.

MEF and Patski mixtures. The same simulation set up as used in the two-sample human cell line simulation is applied to 5 libraries (ML1, ML2, ML3, PL1, and PL2) from the Ramani dataset to mix the two mouse embryonic fibroblast cell lines, Patski and MEF.[9] The total number of cells in the simulated mixture matrices with an equal number of cells across

samples is 918. The total number of cells in the reduced mixtures for the simulated mixture matrices where the total number of cells are different is 460.

Computation Test with 10Kb Hi-C data.

In order to assess the computational speed of *THUNDER* on genome-wide Hi-C data, we apply *THUNDER* to intrachromosomal Hi-C data at 10Kb resolution in YRI samples.[7] We randomly select 5 samples to be included in the analyses. First, we perform feature selection for each chromosome through simple parallelization. Then, we concatenate the subset mixture matrices as input for the final deconvolution estimate. We use computing time and memory usage to assess the computational efficiency for both feature selection and estimation of cell type proportions across the three datasets. *THUNDER* feature selection was computed allowing 2 days of computing time, 128GB of available memory and NMF with random initialization ran for 100 iterations.

References

1. Shen-Orr, S. S. *et al.* Cell type-specific gene expression differences in complex tissues. *Nat. Methods* (2010). doi:10.1038/nmeth.1439
2. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* (2015). doi:10.1038/nmeth.3337
3. Li, Z. & Wu, H. TOAST: improving reference-free cell composition estimation by cross-cell type differential analysis. *Genome Biol.* **20**, 1–17 (2019).

4. Zheng, S. C., Breeze, C. E., Beck, S. & Teschendorff, A. E. Identification of differentially methylated cell types in epigenome-wide association studies. *Nat. Methods* (2018). doi:10.1038/s41592-018-0213-x

5. Houseman, E. A. *et al.* Reference-free deconvolution of DNA methylation data and mediation by cell composition effects. *BMC Bioinformatics* (2016). doi:10.1186/s12859-016-1140-4

6. Jaffe, A. E. & Irizarry, R. A. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol.* (2014). doi:10.1186/gb-2014-15-2-r31

7. Gorkin, D. U. *et al.* Common DNA sequence variation influences 3-dimensional conformation of the human genome. *Genome Biol.* (2019). doi:10.1186/s13059-019-1855-4

8. Tan, L., Xing, D., Chang, C. H., Li, H. & Xie, X. S. Three-dimensional genome structures of single diploid human cells. *Science (80-.).* (2018). doi:10.1126/science.aat5641

9. Ramani, V. *et al.* Massively multiplex single-cell Hi-C. *Nat. Methods* (2017). doi:10.1038/nmeth.4155

10. Stevens, T. J. *et al.* 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature* (2017). doi:10.1038/nature21429

11. Good, I. J. & Mittal, Y. The Amalgamation and Geometry of Two-by-Two Contingency Tables. *Ann. Stat.* (1987). doi:10.1214/aos/1176350369

12. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* (2012). doi:10.1038/nature11082

13. Schmitt, A. D. *et al.* A Compendium of Chromatin Contact Maps Reveals Spatially

Active Regions in the Human Genome. *Cell Rep.* (2016). doi:10.1016/j.celrep.2016.10.061

14. Crowley, C. *et al.* FIREcaller: an R package for detecting frequently interacting regions from Hi-C data. *bioRxiv* (2019). doi:10.1101/619288
15. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* (2014). doi:10.1016/j.cell.2014.11.021
16. Xu, Z., Zhang, G., Wu, C., Li, Y. & Hu, M. FastHiC: A fast and accurate algorithm to detect long-range chromosomal interactions from Hi-C data. in *Bioinformatics* (2016). doi:10.1093/bioinformatics/btw240
17. Xu, Z. *et al.* A hidden Markov random field-based Bayesian method for the detection of long-range chromosomal interactions in Hi-C data. *Bioinformatics* (2016). doi:10.1093/bioinformatics/btv650
18. Sefer, E., Duggal, G. & Kingsford, C. Deconvolution of ensemble chromatin interaction data reveals the latent mixing structures in cell subpopulations. *J. Comput. Biol.* (2016). doi:10.1089/cmb.2015.0210
19. Junier, I., Spill, Y. G., Marti-Renom, M. A., Beato, M. & Le Dily, F. On the demultiplexing of chromosome capture conformation data. *FEBS Letters* (2015). doi:10.1016/j.febslet.2015.05.049
20. Carstens, S., Nilges, M. & Habeck, M. Bayesian inference of chromatin structure ensembles from population-averaged contact data. *Proc. Natl. Acad. Sci. U. S. A.* (2020). doi:10.1073/pnas.1910364117
21. Martin, J. S. *et al.* HUGIn: Hi-C unifying genomic interrogator. *Bioinformatics* (2017). doi:10.1093/bioinformatics/btx359

22. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* (2019). doi:10.1093/nar/gky1120

23. Li, Y., Hu, M. & Shen, Y. Gene regulation in the 3D genome. *Human molecular genetics* (2018). doi:10.1093/hmg/ddy164

24. Mumbach, M. *et al.* HiChIP: Efficient and sensitive analysis of protein-directed genome architecture. *bioRxiv* (2016). doi:10.1101/073619

25. Fang, R. *et al.* Mapping of long-range chromatin interactions by proximity ligation-assisted ChIP-seq. *Cell Research* (2016). doi:10.1038/cr.2016.137

26. Juric, I. *et al.* Maps: Model-based analysis of long-range chromatin interactions from PLAC-seq and HiChIP experiments. *PLoS Comput. Biol.* (2019). doi:10.1371/journal.pcbi.1006982

27. Song, M. *et al.* Cell-type-specific 3D epigenomes in the developing human cortex. *Nature* (2020). doi:10.1038/s41586-020-2825-4

28. Rowley, M. J. & Corces, V. G. Organizational principles of 3D genome architecture. *Nat. Rev. Genet.* (2018). doi:10.1038/s41576-018-0060-8

29. Kim, P. M. & Tidor, B. Subsystem identification through dimensionality reduction of large-scale gene expression data. *Genome Res.* (2003). doi:10.1101/gr.903503

30. Brunet, J. P., Tamayo, P., Golub, T. R. & Mesirov, J. P. Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl. Acad. Sci. U. S. A.* (2004). doi:10.1073/pnas.0308531101

31. Pehkonen, P., Wong, G. & Törönen, P. Theme discovery from gene lists for identification and viewing of multiple functional groups. *BMC Bioinformatics* (2005).

doi:10.1186/1471-2105-6-162

32. Gaujoux, R. & Seoighe, C. A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics* (2010). doi:10.1186/1471-2105-11-367

33. Lee, D. D. & Seung, H. S. Algorithms for non-negative matrix factorization. in *Advances in Neural Information Processing Systems* (2001).

34. Stanford University. CIBERSORT. (2020).