THE UNIVERSITY OF CHICAGO

# MODEL BASED VISUALIZATION OF STRUCTURE IN BIOLOGICAL DATA

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS

BY
KUSHAL K. DEY

CHICAGO, ILLINOIS
AUGUST 2018

Dedicated to my first Statistics mentor, Prof. P.K. Giri

# Table of Contents

List of Figures . . . . . vi

List of Tables . . . . . xiv

Acknowledgments . . . . . xv

Abstract . . . . . xvi

1 Introduction . . . . . 1
- 1.1 Preface . . . . . 1
- 1.2 Overview . . . . . 2
  - 1.2.1 Chapter 2 . . . . . 2
  - 1.2.2 Chapter 3 . . . . . 3
  - 1.2.3 Chapter 4 . . . . . 3
  - 1.2.4 Chapter 5 . . . . . 4

2 *CountClust*: Clustering and visualization of structure in RNA-seq data using a Grade of Membership Model . . . . . 6
- 2.1 Introduction . . . . . 6
- 2.2 Methods Overview . . . . . 8
- 2.3 Results . . . . . 9
  - 2.3.1 Bulk RNA-seq data of human tissue samples . . . . . 9
  - 2.3.2 Sub-analysis of Brain tissues . . . . . 12
  - 2.3.3 Single-cell RNA-seq data . . . . . 13
- 2.4 Discussion . . . . . 19
- 2.5 Methods and Materials . . . . . 22
  - 2.5.1 Model Fitting . . . . . 22
  - 2.5.2 Visualizing Results . . . . . 22
  - 2.5.3 Cluster annotation . . . . . 23
  - 2.5.4 Comparison with hierarchical clustering . . . . . 23
  - 2.5.5 Thinning . . . . . 24
  - 2.5.6 Code Availability . . . . . 24
- 2.6 Author contributions . . . . . 25

3 *aRchaic* : Modeling and Visualization of DNA damage patterns using a Grade of Membership Model . . . . . 33
- 3.1 Introduction . . . . . 33
- 3.2 Methods . . . . . 35
- 3.3 Results . . . . . 37
  - 3.3.1 `aRchaic` clustering of modern and ancient individuals . . . . . 37
  - 3.3.2 The effects of contamination on inferred grades of membership . . . . . 38
  - 3.3.3 `aRchaic` can identify both DNA damage and technical artifacts . . . . . 39

3.4 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 40
3.5 Author Contributions . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 41
3.6 Figures . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 41

4 *Logolas* : Enrichment Depletion and String Logo Plots . . . . . . . . . . . . . . 46
4.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 46
4.2 Implementation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 47
4.2.1 Intuition . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 47
4.2.2 The *EDLogo* plot . . . . . . . . . . . . . . . . . . . . . . . . . . . . 48
4.2.3 Interpretation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 51
4.2.4 A variation: the scaled *EDLogo* plot . . . . . . . . . . . . . . . . . 51
4.3 Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 52
4.3.1 Comparison with existing logo plots . . . . . . . . . . . . . . . . . . 52
4.3.2 The *scaled EDLogo* plot . . . . . . . . . . . . . . . . . . . . . . . . 55
4.3.3 Further Variations . . . . . . . . . . . . . . . . . . . . . . . . . . . . 56
4.4 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 56
4.5 Author contributions . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 56
4.6 Supplementary Methods . . . . . . . . . . . . . . . . . . . . . . . . . . . 57

5 *CorShrink*: Adaptively Parsimonious Representation of Correlation Matrices . . . 59
5.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 59
5.2 Methods . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 60
5.3 Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 63
5.3.1 Applications - Genetics . . . . . . . . . . . . . . . . . . . . . . . . . 63
5.3.2 Simulation studies . . . . . . . . . . . . . . . . . . . . . . . . . . . . 65
5.3.3 Applications - Natural Language Processing . . . . . . . . . . . . . . 67
5.4 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 68
5.5 Author contributions . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 69

A Supplementary Figures . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 74

References . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 102

# List of Figures

2.1 **(a)**: Structure plot of estimated membership proportions for GoM model with $K = 20$ clusters fit to 8555 tissue samples from 53 tissues in GTEx data. Each horizontal bar shows the cluster membership proportions for a single sample, ordered so that samples from the same tissue are adjacent to one another.**(b)**: Structure plot of estimated membership proportions for $K = 4$ clusters fit to only the brain tissue samples. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 29

2.2 Visualization of the same GTEx data as in Figure 1 (a) across all tissues using standard and widely used approaches - Principal Component Analysis (PCA), Multi dimensional Scaling (MDS), t-SNE and hierarchical clustering. All the analysis are done on log CPM normalized expression data to remove library size effects. **(a)**: Plot of PC1 vs PC2 on the log CPM expression data, **(b)**: Plot of first two dimensions of the t-SNE plot, **(c)** Plot of first two dimensions of the Multi-Dimensional Scaling (MDS) plot. **(d)** Dendrogram for the hierarchical clustring of the GTEx tissue samples based on the log CPM expression data. . . 30

2.3 Structure plot of estimated membership proportions for GoM model with $K = 7$ clusters fit to $1,041$ single cells from [57]. The samples (cells) are ordered so that samples from the same amplification batch are adjacent and within each batch, the samples are sorted by the proportional representation of the underlying clusters. In this analysis the samples do not appear to form clearly-defined clusters, with each sample being allocated membership in several "clusters". Membership proportions are correlated with batch, and some groups of batches (e.g. 28-29; 32-45) show similar palettes. These results suggest that batch effects are likely influencing the inferred structure in these data. . . . . . . . . . . . . . . . . . . . 31

2.4 Structure plot of estimated membership proportions for GoM model with $K = 6$ clusters fit to 259 single cells from [22]. Each cluster is annotated by the genes that are most distinctively expressed in that cluster, and by the gene ontology categories for which these distinctive genes are most enriched. . . . . . . . . . . 32

3.1 Illustration of the `aRchaic` grades of membership and mismatch profiles. (a) The features of a mismatch modeled by `aRchaic` (b) A depiction of an ancient DNA sample that has 80% of it's reads assigned to cluster 1 cluster and 20% of it's reads assigned to cluster 2. Each cluster is defined by a *mismatch profile* showing the enrichment of the mismatch type, bases flanking the mismatch, the distance of the mismatch from the nearest end of the read, and the base immediately 5' to the strand-break. To produce a mismatch profile for a cluster, mismatch features are aggregated across reads assigned to the cluster, and their frequencies are represented by an *EDLogo* plot [26]. In the *EDLogo* plot, the frequencies are scaled against a background frequency computed from The 1000 Genomes Project [1]. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 42

3.2 aRchaic clearly distinguishes between modern, ancient (UDG), and ancient (non-UDG) samples aRchaic is applied with $K = 3$ to a collection of ancient individuals from four studies [42, 68, 79, 116] along with modern individuals randomly sampled from the 1000 Genomes Project and the Human Genome Diversity Panel [1, 17]. Modern samples have high membership in the red cluster. The *EDLogo* representation of this cluster does not show strong enrichment against a modern background. The ancient (non-UDG) samples are representative of the blue cluster. The *EDLogo* plot for the blue cluster shows a strong enrichment in $C \to T$ mismatches at the end of reads, a depletion of Guanine in the right flanking base, and a depletion of Cytosine at the 5' strand-break. The ancient (UDG) samples have partial membership both in the red cluster and in the gold cluster. The *EDLogo* plot for the gold cluster is enriched in $C \to T$ mismatches at the terminal ends of the reads, and also shows an enrichment of Guanine at the right flanking base. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 43

3.3 Estimated grades of membership reflect levels of contamination (a) Reads from one ancient individual (KO1 from [42]) were split into 10 equally sized groups. To each of these groups, reads (with at least one mismatch) were added from a distinct individual in the 1000 Genomes Project [1] to each group (S1-S10) at varying levels of contamination. (b) We applied aRchaic with $K = 2$ on a combined dataset comprised of these 10 contaminated groups of reads (S1-S10) along with 40 other modern individuals from 1000 Genomes (c) The grades of membership in cluster (red) were plotted as a function of the percentage of contamination. (d) Each group (S1-S10) was further sub-sampled to 10,000 reads, and aRchaic was applied with $K = 2$ to the new subsampled groups and the same 40 modern individuals as in panel b. . . . . . . . . . . . . . . . . . . . . . . . 44

3.4 DNA damage and library preparation techniques drive grades of membership (a) We applied aRchaic with $K = 2$ to 25 modern samples from [73]. The samples prepared with the True-Seq kit show nearly full membership in the pink cluster. Samples prepared with the Nextera kit show partial membership in the pink cluster and the tan cluster. The tan cluster shows a blip at the 12th position from the end of the read (b) We applied aRchaic with $K = 2$ to 25 ancient samples from [73]. The two clusters show an enrichment of $C \to T$ mismatches at the ends of reads and an enrichment of purines at the 5' strand-break. . . . . 45

4.1 **Illustration of the differences between standard logo, *EDLogo* and wKL-Logo representations.** The figure shows how the different logos represent observed frequencies $\mathbf{p} = (p_A, p_C, p_G, p_T) = (0.33, 0.33, 0.33, 0.01)$ (compared with a uniform background, $\mathbf{q} = (0.25, 0.25, 0.25, 0.25)$). The standard logo effectively represents $\mathbf{p}$ by highlighting that "A, C and G are enriched"; *EDLogo* represents it by highlighting "T is depleted"; wKL-Logo represents it as "A, C and G are enriched and T is depleted". All are correct statements, but the *EDLogo* representation is the most parsimonious. . . . . . . . . . . . . . . . . 48

4.2 **Comparison of standard logo plot, weighted KL (w-KL) logo plot and *EDLogo* plot on four examples.** Panel (a): the transcription factor binding site of the EBF1-disc1 transcription factor. Panel (b): the binding motif (Motif2 Start=257 Length=11) of the protein *D-isomer specific 2-hydroxyacid dehydrogenase, catalytic domain (IPR006139)* from [61, 111]. Panel (c): mutational signature profile of mutations in lymphoma B cells, with data from [4]. The depletion of G to the right of the mutation - possibly occurring due to the rarity of CpG sites owing to de-amination of methylated cytosines - is clearest in the *EDLogo* representation. Panel (d): relative abundance of histone modification sites across various genomic regions in the lymphoblastoid cell line GM06990 (Table S2 in Koch et al 2007 [64]). These examples illustrate the ability of the *EDLogo* plot to highlight both enrichment and depletion, while avoiding unnecessarily visual clutter. The last two examples also illustrate how our software allows arbitrary strings as elements in a logo plot. . . . . . . . . . . . . . . . . . . . . . . . . 53

5.1 (a) The image plot of the pairwise correlation matrix between tissue pairs for the log CPM expression data of the *PLIN1* gene. (b) The probability and cumulative density function plots for the empirically estimated mixture of half-uniform prior used to shrink the correlations in `CorShrink`. The black dot represents the prior probability mass of observing 0 correlation. (c) The image plot the estimated correlation matrix due to `CorShrink`. The representation is visually more parsimonious and arguably more interpretable than (a). (d) plots the pairwise sample correlation values against the `CorShrink` fitted estimates for each tissue pair in a scatter plot and colors each point based on the number of matched samples for the corresponding tissue pair. Expectedly the pairs with low numbers of matched samples (light colored points) undergo high shrinkage while those with larger number of matched samples remain largely unperturbed by `CorShrink`. . 70

5.2 Box plot of the Correlation Matrix Distance (CMD) [50] between population correlation matrix and the estimated matrix from different methods - *corpcor*[106, 107], `CorShrink`, *PDSCE* [101], GLASSO [40] at different tuning parameters and the empirical pairwise correlation matrix, for different structural assumptions on the underlying population correlation - Hub structure, Toeplitz structure and a banded precision matrix, see Supplementary Figure S27. `CorShrink` outperforms the other methods for the structured/sparse covariance models (Hub and Toeplitz), with *PDSCE* being closest to `CorShrink` in performance. . . . . . . . 71

5.3 Plots of sorted square-root eigenvalue trends of the population correlation matrix against those of estimated correlation matrices using different methods - *corpcor* [106, 107], `CorShrink`, *PDSCE* [101], GLASSO [40] at different tuning parameters and the empirical pairwise correlation matrix, for different structural assumptions on the underlying population correlation - hub structure, Toeplitz structure and a banded precision matrix structure, see Supplementary Figure S27. The trend of sorted eigenvalues for `CorShrink` follow that of the original matrix closely for the Hub and the Toeplitz models. . . . . . . . . . . . . . . . . . . . . . . . . 72

5.4 We extracted the top 1000 contextually close words to each of the two word sets of our interest - $\{$*martin, luther, king*$\}$ and $\{$*civil, rights*$\}$ based on *word2vec* analysis of the monthly issues of the Ebony magazine in 1968 and combined these two sets of words. For each pair of words in the combined word set, we plotted the cosine similarities before and after the `CorShrink` adjustment, colored by how many times they occurred in the texts. We report the top 25 words contextually close to the word sets of interest before and after the `CorShrink` adjustment. The `CorShrink` adjustment seems to remove terms that are specific to a few texts or specific to certain events and instead incorporate more broadly similar words to our word sets of interest in the top 25 lists. Examples are discussed in depth in the Results section. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 73

S1 **Structure plot of GTEx V6 tissue samples for (a) $K = 5$, (b) $K = 10$, (c) $K = 15$, (d) $K = 20$.** Some tissues form a separate cluster from the other tissues from $K = 5$ onwards (for example: Whole Blood, Skin), whereas some tissue only form a distinctive subgroup at $K = 20$ (for example: Arteries). . . . 74

S2 **Top five principal components (PC) for GTEx V6 tissue samples.** Scatter plot representation of the top five PCs of the GTEx tissue samples. Data was transformed to log2 counts per million (CPM). . . . . . . . . . . . . . . . . . . . 75

S3 **Comparison between GoM model and hierarchical clustering under different scenarios of data transformation.** We used GTEx V6 data for model performance comparisons. Specifically, for every pair of the 53 tissues, we assessed the ability of the methods to separate samples according to their tissue of origin. The subplots of heatmaps show the results of evaluation under different scenarios. Filled squares in the heatmap indicate successful separation of the samples in corresponding tissue pair comparison. (a) Hierarchical clustering on log2 counts per million (CPM) transformed data using Euclidean distance. (b) Hierarchical clustering on the standardized log2-CPM transformed data (transformed values for each gene was mean and scale transformed) using the Euclidean distance. (c) GoM model of $K = 2$ applied to counts. (d) Hierarchical clustering on counts data with the assumption that, for each gene the sample read count $c_{ng}$ has a variance $\bar{c}_g + 1$. (e) Hierarchical clustering applied to adjusted count data. Each gene has a mean expression value of 0 and variance of 1. The GoM model with $K = 2$ is able to separate samples of different tissue of origin, better than hierarchical cluster methods. . . . . . . . . . . . . . . . . . . . . . . . . . . 76

S4 **GTEx brain PCA, t-SNE and MDS.** . . . . . . . . . . . . . . . . . . . . . 77

S5 **GTEx brain PCA, t-SNE and MDS.** . . . . . . . . . . . . . . . . . . . . . 78

S6 **A comparison of accuracy of hierarchical clustering vs GoM on thinned GTEx data, with thinning parameters of $p_{thin} = 0.01$ and $p_{thin} = 0.001$.** For each pair of tissue samples from the GTEx V6 data we assessed whether or not each clustering method (with $K = 2$ clusters) separated the samples according to their tissue of origin, with successful separation indicated by a filled square. Thinning deteriorates accuracy compared with the unthinned data (Fig 2), but even then the model-based method remains more successful than the hierarchical clustering in separating the samples by tissue or origin . . . . . . . . . . . . . . 79

S7 **Deng et al (2014) PCA, tSNE, MDS and dendrogram plots for hierarchical clustering.** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 80

S8 **Additional GoM analysis of Deng et al (2014) data including blastocyst samples and 48 blastocyst marker genes.** We considered 48 blastocyst marker genes (as chosen by Guo et al., 2010) and fitted GoM model with $K = 3$ to 133 blastocyst samples. In the Structure plot, blastocyst samples are arranged in order of estimated membership proportion in the Green cluster. The panel located above the Structure plot shows the corresponding pre-implantation stage from which blastocyst samples were collected. The heatmap located below the Structure plot represents expression levels of the 48 blastocyst marker genes (log2 CPM), and the corresponding dendrogram shows results of hierarchical clustering (complete linkage). The table on the right of the expression heatmap displays gene information, showing, from left to right, 1) whether or not the gene is a transcription factor, 2) the driving GoM cluster if the gene was among the top five driving genes, and 3) the featured cell type (TE: trophecoderm, EPI: epiblast, PE: primitive endoderm) that was found in Guo et al., 2010. . . . . . . . . . . . 81

S9 **Sparse Factor Analysis loadings visualization of GTEx V6 tissue samples.** The colors represent the 20 different factors. The factor loadings are presented in a stacked bar for each sample. We performed SFA under the scenarios of when the loadings are sparse (left panel) and when the factors are sparse (right panel). . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 82

S10 **Sparse Factor Analysis loadings visualization of GTEx brain tissue samples.** The colors represent the 6 different factors. The factor loadings are presented in a stacked bar for each sample. We performed SFA under the scenarios of when the loadings are sparse (left panel) and when the factors are sparse (right panel). . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 83

S11 **Sparse Factor Analysis loadings visualization of mouse pre-implantation embryos from Deng et al., (2014). The colors represent the 6 different factors. The factor loadings are presented in a stacked bar for each sample. We performed SFA under the scenarios of when the loadings are sparse (left panel) and when the factors are sparse (right panel).** . . 84

S12 `aRchaic` grades of membership for the example in Fig 4.2 corresponding to 3 different values of $K$ ($K = 4, 5, 6$). Higher values of $K$ distinguish among the ancient studies, reflecting lab and study specific biases. . . . . . . . . . . . . . 85

S13 We apply aRchaic with $K = 6$ on the data from Fig 3.4. In addition to separating out the ancients from the moderns, aRchaic now distinguishes between moderns individuals based on library kit (Nextera vs Tru-seq.) . . . . . . . . . . . . . . . . 86

S14 aRchaic plot for $K = 2$ on the combined data of 25 moderns and 25 ancients from [73]. aRchaic clearly distinguishes the moderns from the ancients. The ancients are primarily presented by the blue cluster. This cluster shows an enrichment of $C \rightarrow T$ mismatches and depletion of $T \rightarrow C$ mismatches with respect to modern background, as well as enrichment of G and depletion of T at the 5' strand break. The red cluster shows a blip at 12th position from the end of the read, the explanation for which is provided S15. . . . . . . . . . . . . . . . . . . 87

S15 The frequency of all mismatch types plotted against the position of the read (from the 5' end) for each of the 25 moderns samples in [73]. Each sample was prepared by one of two library kits: Nextera and True-Seq. Most the samples prepared with the Nextera kit show a spike in frequency at the 12th position from the 5' end of the read . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 88

S16 **Illustration of "mirror property" of *EDLogo*.** *Panel (a)*: *EDLogo* plot of the position weight matrix (PWM) of the primary discovered motif *disc1* from [63] of the EBF1 transcription factor against uniform background. *Panel (b)*:*EDLogo* plot of a uniform PWM against the PWM of EBF1 as background. That is, panels (a) and (b) are comparing the same two PWMs, but differ in which one they treat as the "background". The *EDLogo* plot obeys the mirror property, in that (b) is a mirror image of (a) (modulo the orientation of the symbols, which are translated and not reflected). . . . . . . . . . . . . . . . . . . . . . . . . 89

S17 ***EDLogo* plots for six different motifs of the EBF1 transcription factor.** The PWMS for *known1* and *known2* come from the TRANSFAC database [135]; *known3* from the JASPAR database [103]; *known4* from [59]; *disc1* and *disc2* were discovered by the ENCODE project [63]. Three of the motifs (*known3*, *known4* and *disc1*) show depletion of G and C in the middle of the binding site. 90

S18 **Comparison of the *EDLogo* plot (a) with *pmsignature* [115] plot (b) for visualizing cancer mutational signatures.** Both plots show a signature of lymphoma B cell from [4]. The *EDLogo* plot highlights the depletion of $G$ at the right flanking base more clearly than does the *pmsignature* plot. The use of strings to represent mutations in the center is arguably more intuitive than the *pmsignature* representation. . . . . . . . . . . . . . . . . . . . . . . . . . . . 91

S19 *EDLogo* plots for the mutation signature profiles of 24 different cancer types from [4]. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 92

S20 **Illustration of *scaled EDLogo* plot on examples from Figure 4.2.** The standard logo and unscaled *EDLogo* plots are repeated here to ease comparisons. The *scaled EDLogo* plot highlights strong enrichments more than the unscaled version and may be preferred in settings when enrichments are the primary focus. 93

S21 **Illustration of various options for *EDLogo* plot.** Each plot shows an *EDLogo* plot for a specific binding motif (Motif2 Start=257 Length=11) of the protein *D-isomer specific 2-hydroxyacid dehydrogenase, catalytic domain (IPR006139)* against a uniform background. The plots illustrate the use of several different scoring schemes (*log ratio*, *log odds ratio*, *ratio* and *probKL*) with and without scaling by the symmetric Kullback-Leibler divergence. See Supplementary Methods for details on the scoring schemes. (Note that only the *log ratio* and *log odds ratio* scoring schemes satisfy the "mirror property".) . . . . . . . . . . . . . . . 94

S22 **Illustration of median adjustment of a position specific scoring matrix (PSSM).** The PSSM shown here is for the binding motif of the protein *D-isomer specific 2-hydroxyacid dehydrogenase, catalytic domain (IPR006139)* (Motif2,Start=257, Length=11). The median adjusted PSSM Logo (*bottom panel*) is arguably less cluttered than the non-adjusted version (*top panel*). . . . . . . . . 95

S23 Application of `CorShrink` on the donor by tissue expression data for the *PLIN1* gene using different mixture model priors than the half-uniform mixture prior used in Figure 5.1. (a) presents the image plot for the pairwise sample correlation plot, (b) presents the `CorShrink` model fit with mixture normal prior centered around 0, (c) presents the `CorShrink` model fit with mixture normal prior centered around an estimated mode, (d) presents the `CorShrink` model fit using an essentially non-parametric prior (see Methods for definition of these models). . . 96

S24 Image plots of the estimated correlation matrices using (a) gene-wide `CorShrink` and (b) tissue-wide `CorShrink`, both with half-uniform mixture model prior on the subject by tissue expression matrix data for the *PLIN1* gene. Both representations are visually broadly equivalent. . . . . . . . . . . . . . . . . . . . . . . . 97

S25 Image plots of the estimated correlation matrices using both tissue-wide `CorShrink` and gene-wide `CorShrink`, using half-uniform mixture model prior, on the subject by tissue expression data for three different genes - *HBB*, *MTURN* and *VSIR*. *HBB* correlation patterns are similar to *PLIN1* in Figure 5.1 with high correlation among the Brain tissues and negligible correlation among other tissues. *MTURN* exhibits low correlation across all tissues and *VSIR* exhibits high correlation across almost all the tissues. . . . . . . . . . . . . . . . . . . . . . 98

S26 Comparison of the tissue-tissue correlation matrix of the log CPM expression data for *PLIN1* gene (a) and the estimated correlation matrix from our proposed `CorShrink` method (b) with respect to correlation matrices obtained after imputing the missing observations in the data matrix by SoftImpute (c) and FLASH (d). `CorShrink` seems to capture the subtle structure in the tissue tissue correlations and generate a visually more interpretable representation than the imputation mechanisms. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 99

S27 A demonstration of the population correlation and inverse correlation structure from which simulation studies were carried out. . . . . . . . . . . . . . . . . . . 100

S28 Box plot of the Frobenius distance between population correlation matrix and the estimated matrix from different methods - *corpcor* [106, 107],`CorShrink`, *PDSCE* [101], GLASSO [40] at different tuning parameters and the empirical pairwise correlation matrix, for different structural assumptions on the underlying population correlation - Hub structure, Toeplitz structure and a banded precision matrix, see Supplementary Figure S27. `CorShrink` outperforms the other methods for the structured/sparse covariance models (Hub and Toeplitz), with *PDSCE* being closest to `CorShrink` in performance.. . . . . . . . . . . . . . . . . . . . . . 101

# List of Tables

2.1 **Cluster Annotations GTEx V6 data (with GO annotations).** . . . . . . . 26

2.2 **Cluster Annotations for GTEx V6 Brain data.** . . . . . . . . . . . . . . . . . 27

2.3 **Cluster Annotations for Deng et al (2014) data.** . . . . . . . . . . . . . . . 28

## Acknowledgments

First and foremost, I would like to thank my advisor, Prof. Matthew Stephens, for his constant support and encouragement. There is a lot to learn from Matthew when it comes to statistics and statistical genetics, but even more to learn from him as a person and from his outlook towards research. Working with him has substantially changed my perspective towards research and I am grateful to receive his mentorship.

I also thank the other members of my thesis committee, Prof. Dan Nicolae and Prof. John Novembre for providing valuable feedback and guidance at every juncture of this project.

I wholeheartedly thank people from Stephens Lab, Novembre Lab, He Lab and Di Rienzo Lab for stimulating discussions and helpful feedback ad support to my research - I am indebted to them for much of my understanding of biology, computation and statistics. I thank my collaborators in different projects - Gao Wang, Hussein Al-Asadi, Chiaowen Joyce Hsiao, Alexander White, Kevin Luo, Sebastian Pott, Lei Sun, Lauren M. Jackson, Dongyue Xie - it was special working with each one of you. I thank Peter Carbonetto, Nan Xiao and John Blischak for always being patient with my frequent computation related problems.

I thank my parents for providing unfailing support to my academic career.

Lastly, I thank my friends and colleagues - Ankita Naha, Moumanti Poddar, Sayar Karmakar, Soudeep Deb, Rounak Dey, Vishwas Srivastava, Swati Ramachandran, Krithika Mohan, Nikita Mishra, Somak Dutta, Rishideep Roy, Subhajit Goswami, Sabyasachi Chatterjee, Sarah Urbut, Severin Thaler, Raman Shah, Shailee Mehta - all of you contributed in making this journey of 5 years a wonderful one.

# Abstract

Biological Data comes in varied forms and the scale of the data is typically large, which often necessitates distinct modeling frameworks and tools to process, analyze and visually summarize the data. An overarching theme of this doctoral thesis is to suggest such novel model based visualization tools for different biological problems.

The second chapter of this thesis extends the concept of a mixed membership model, popularly known as ADMIXTURE model in population genetics and *topic model* in Natural Language Processing (NLP), to the context of RNA-sequencing read expression data in genetics. Applied to data from the GTEx project on 53 human tissues, this approach highlights similarities among biologically-related tissues and identifies distinctively-expressed genes that recapitulate known biology. Applied to single-cell expression data from mouse preimplantation embryos, this approach highlights both discrete and continuous variation through early embryonic development stages, and identifies genes involved in a variety of relevant processesfrom germ cell development.

The third chapter extends similar mixed membership models to analyzing DNA damage patterns in ancient DNA (aDNA) samples, and explore and jointly summarize multiple aDNA samples together with modern samples. Applied to a combined data of modern and ancient individuals from multiple studies, this approach clearly distinguished moderns and ancients irrespective of DNA extraction, lab and sequencing protocols. Additionally. we found that the grades of membership from the fitted mixed membership models can be reflective of relative levels of contamination in the data.

The visual summary of DNA damage patterns, depicted above, includes a version of

logo plot that highlights enrichment and depletion of damage features with respect to a background level of mismatch features computed from modern individuals. We call this representation the *Enrichment Depletion Logo (EDLogo)* plot and present a comprehensive overview of this logo plot in the fourth chapter. We also propose an extension of typically character driven logo plots to string based logos.

In the fifth chapter, we propose an adaptive method for shrinking correlation matrices that leads to a parsimonious representation of the underlying association structure between variables. This method is flexible in handling data matrices with missing observations and accounts for the differences in the number of samples with non-missing observations for a pair of variables in a model based way. Even with no missing data, under small n, large p settings, this method outperforms other popular approaches to correlation shrinkage and is flexible enough to extend to other correlation-like quantities such as the word-word cosine similarity values from *word2vec* models in Natural Language Processing (NLP).

# Chapter 1

# Introduction

## 1.1 Preface

Early on in my tenure as a graduate student, I was showing my advisor, Matthew Stephens, a paper draft I had written. The draft was about 30 pages long, with 15 figures. Upon glancing over it, Matthew quipped -

*Can you summarize everything you did here into a single figure?*

.

This basic idea – summarizing a seemingly complex data problem into a single visualization that encompasses all major aspects of the analysis and the data - eventually turned out to be the overarching theme for of all my research topics. My limited experience of working at the interface of statistics and biology has made me realize that visualization is the easiest medium to connect with and start a dialogue with people from different domains. Much of my PhD research is an attempt to mix statistical methods with careful visualization of the results, that will hopefully appeal to biologists, archaeologists or ecologists, among others.

Biological data comes in varied forms, which is why it is an exciting field for a computational person to work in. These data often necessitate development of new methods and tools for analysis and subsequent visualization. Also given the fast moving nature of the field and the scale of the data getting bigger with every passing day, there is growing need to make the methods scalable to large data and to make them available as open source software for fast reproducibility.

## 1.2 Overview

The four chapters of this thesis each focus on developing modeling framework to visually summarize different types of biological data. Each of these methods is also available as open source software packages, together with codes to reproduce the figures in this thesis.

### *1.2.1 Chapter 2*

Chapter 2 extends the concept of a mixed membership model or a Grade of Membership model (GoM) [34], popularly known as ADMIXTURE model in population genetics [89] and topic model in Natural Language Processing [15], to the context of RNA-seq expression data. Bulk RNA-seq data as in Gentotype Tissue Expression (GTEx) Project [19] reports the amount of mRNA (which can be considered as a measure of protein created) transcribed from the protein coding genes (called *mRNA expression* in biology jargon) for a particular sample of a tissue. GTEx in fact has mRNA expression data across all protein coding genes for 8555 such tissue samples spanning across 53 different tissues.

Each tissue sample (say a Blood sample) may comprise of cells of different cell types (say RBC, WBC, platelets) and these cell types are present in varying proportions in the sample under consideration. The main idea of this chapter is to estimate clusters as a potential analog to cell types, that represent a distinct mRNA expression profile and also estimate the proportional representation of these clusters in each tissue sample. We observe that this method indeed identifies biological subgroups driven by marker genes for those subgroups, and also allows for a sparse representation of the tissue samples by the clusters. This method can also be flexibly applied to single cell level data as well, where it can identify latent cell states and the how these cell states contribute to determining the mRNA expression profile for the given cell. For more details about the method and its application in RNA-seq context, see Chapter 2.

### *1.2.2 Chapter 3*

In Chapter 3, we extend the GoM model to visually summarize DNA damage patterns in ancient DNA samples. A major issue in processing ancient DNA (aDNA) data is to control for contamination from modern human DNA. Fortunately, aDNA contains unique signatures that can be used to distinguish it from modern DNA. These signatures are left by DNA-damaging processes that accumulate over time. The predominant signatures of damage are a high frequency of cytosine to thymine substitutions ($C \rightarrow T$) at the ends of fragments, and elevated rates of purines (A & G) before the 5' strand-breaks. To assess DNA damage, a common QC procedure is to plot for each sample, the $C \rightarrow T$ mismatch rate along the read and the composition of bases before the 5' strand-breaks. Though simple and useful, this procedure has several limitations.

We propose a modeling framework based on the GoM model where each sample has a grade of membership in each of the K different damage profiles that are estimated from the data. This has several advantages over existing approaches. Our method flexibly learns the important features, and produces a single concise visual summary of the data. Additionally, through the grades of membership, this approach quantifies the relative degrees of exogenous modern contamination. We applied this method to a combined data-set comprised of ancient samples from several aDNA studies and modern individuals from the 1000 Genomes Project [1]. It clearly distinguished modern and ancient individuals irrespective of DNA extraction, lab and sequencing protocols. Additionally. we found that the estimated grades of membership accurately reflected relative levels of contamination in the data.

### *1.2.3 Chapter 4*

The work presented in Chapter 4 is a spin-off of the DNA damage work in Chapter 3. To effectively represent the cluster from the GoM model fit on the DNA damage data in Chapter

3, we realized that a logo type plot [8, 109] would be the perfect candidate. However there were two primary concerns regarding using logo plots for this problem. Firstly, we needed to present not just letters - say A, C, G and T - as standard logo plotting softwares do, but also strings such as $C \to T, T \to A$ etc to denote mismatches. Secondly, we needed to highlight both enrichment and depletion of features in each cluster with respect to a known modern background for the benefit of comparison. The standard logo plots are more biased in their representation of enrichment of symbols than depletion. We came up with a visualization package called *Logolas* that not only has the flexibility to plot string logos, like the mismatches, but also highlights both enrichment and depletion of symbols. Chapter 4 gives a detailed overview of how these logo plots were computed and how it can be used for a more parsimonious representation of conserved patterns of bases/amino acids in an aligned sequence of DNA, RNA or proteins.

### *1.2.4 Chapter 5*

In Chapter 5, we propose an adaptive strategy to shrink correlation matrices from a data matrix with potentially large scale missing observations. Correlation shrinkage is an extensively studied field in statistics. A large proportion of these studies are focused on efficient estimation under *small n, large p* settings. But not many of these methods can effectively handle missing observations. Our method, built on the adaptive shrinkage (*ash*) framework proposed in [119] is not only competitive with other shrinkage approaches in small n, large p settings, but its main advantage is to produce visually parsimonious representation of correlation structure in an adaptive way for data matrices with high proportion of missing data. As an example application, `CorShrink` is applied to the donor by tissue expression data matrix for a gene in the Genotype Tissue Expression (GTEx) project [19], where the data contains large scale missing observations owing to each donor contributing only a few tissues. The estimated correlation matrix using `CorShrink` is found to be less visually cluttered and more interpretable than the corresponding pairwise sample correlation matrix.

# Chapter 2

# *CountClust*: Clustering and visualization of structure in RNA-seq data using a Grade of Membership Model

*(with C.J. Hsiao and M. Stephens)*

## 2.1 Introduction

Ever since large-scale gene expression measurements have been possible, clustering – of both genes and samples – has played a major role in their analysis [5, 32, 47]. For example, clustering of genes can identify genes that are working together or are co-regulated, and clustering of samples is useful for quality control as well as identifying biologically-distinct subgroups. A wide range of clustering methods have therefore been employed in this context, including distance-based hierarchical clustering, $k$-means clustering, and self-organizing maps (SOMs); see for example [27, 58] for reviews.

Here we focus on cluster analysis of samples, rather than clustering of genes (although our methods do highlight sets of genes that distinguish each cluster). Traditional clustering methods for this problem attempt to partition samples into distinct groups that show "similar" expression patterns. While partitioning samples in this way has intuitive appeal, it seems likely that the structure of a typical gene expression data set will be too complex to be fully captured by such a partitioning. Motivated by this, here we analyse expression data using grade of membership (GoM) models [34], which generalize clustering models to allow each sample to have partial membership in multiple clusters. That is, they allow that each sample has a proportion, or "grade" of membership in each cluster. Such models are widely used in population genetics to model admixture, where individuals can have ancestry from multiple populations [89], and in document clustering [14, 15] where each document can have membership in multiple topics. In these fields GoM models are often known as "admixture models", and "topic models" or "Latent Dirichlet Allocation" [15]. GoM models have also recently been applied to detect mutation signatures in cancer samples [115].

Although we are not the first to apply GoM-like models to gene expression data, previous applications have been primarily motivated by a specific goal, "cell type deconvolution", which involves using cell-type-specific expression profiles of marker genes to estimate the proportions of different cell types in a mixture [2, 74, 93]. Specifically, the GoM model we use here is analogous to – although different in detail from – blind deconvolution approaches [96, 110, 132] which estimate cell type proportions and cell type signatures jointly (see also [92, 114] for semi-supervised approaches). Our goal here is to demonstrate that GoM models can be useful much more broadly for understanding structure in RNA-seq data – not only to deconvolve mixtures of cell types. For example, in our analysis of human tissue samples from the GTEx project below, the GoM model usefully captures biological heterogeneity among samples even though the inferred grades of membership are unlikely to correspond precisely to proportions of specific cell types. And in our analyses of single-cell expression data the GoM model highlights interesting structure, even though interpreting the grades of membership as "proportions of cell types" is clearly inappropriate because each sample is a single cell! Here we are exploiting the GoM as a flexible extension of traditional cluster models, which can capture "continuous" variation among cells as well as the more "discrete" variation captured by cluster models. Indeed, the extent to which variation among cells can be described in terms of discrete clusters versus more continuous populations is a fundamental question that, when combined with appropriate single-cell RNA-seq data, the GoM models used here may ultimately help address.

## 2.2 Methods Overview

We assume that the RNA-seq data on $N$ samples has been summarized by a table of counts $C_{N \times G} = (c_{ng})$, where $c_{ng}$ is the number of reads from sample $n$ mapped to gene $g$ (or other unit, such as transcript or exon) [86]. The GoM model is a generalization of a cluster model, which allows that each sample has some proportion ("grade") of membership, in each cluster. For RNA-seq data this corresponds to assuming that each sample $n$ has some proportion of its reads, $q_{nk}$ coming from cluster $k$. In addition, each cluster $k$ is characterized by a probability vector, $\theta_{k\cdot}$, whose $g$th element represents the relative expression of gene $g$ in cluster $k$. The GoM model is then

$$(c_{n1}, c_{n2}, \cdots, c_{nG}) \sim \text{Multinomial}\left(c_{n+}, p_{n1}, p_{n2}, \cdots, p_{nG}\right), \tag{2.1}$$

where

$$p_{ng} := \sum_{k=1}^{K} q_{nk} \theta_{kg}. \tag{2.2}$$

The number of clusters $K$ is set by the analyst, and it can be helpful to explore multiple values of $K$ (see Discussion).

To fit this model to RNA-seq data, we exploit the fact that this GoM model is commonly used for document clustering [15]. This is because, just as RNA-seq samples can be summarized by counts of reads mapping to each possible gene in the genome, document data can be summarized by counts of each possible word in a dictionary. Recognizing this allows existing methods and software for document clustering to be applied directly to RNA-seq data. Here we use the R package `maptpx` [121] to fit the GoM model.

Fitting the GoM model results in estimated membership proportions $q$ for each sample, and estimated expression values $\theta$ for each cluster. We visualize the membership proportions for each sample using a "Structure plot" [99], which is named for its widespread use in visualizing the results of the *Structure* software [89] in population genetics. The Structure plot represents the estimated membership proportions of each sample as a stacked barchart, with bars of different colors representing different clusters. Consequently, samples that have similar membership proportions have similar amounts of each color. See Fig 2.1 for example.

To help biologically interpret the clusters inferred by the GoM model we also implemented methods to identify, for each cluster, which genes are most distinctively differentially expressed in that cluster; that is, which genes show the biggest difference in expression compared with the other most similar cluster (see Methods). Functions for fitting the GoM model, plotting the structure plots, and identifying the distinctive ("driving") genes in each cluster, are included in our R package `CountClust` [25] available through Bioconductor [43].

## 2.3 Results

### *2.3.1 Bulk RNA-seq data of human tissue samples*

We begin by illustrating the GoM model on bulk RNA expression measurements from the GTEx project (V6 dbGaP accession phs000424.v6.p1, release date: Oct 19, 2015, `http://www.gtexportal.org/home/`). These data consist of per-gene read counts from RNA-seq performed on $8,555$ samples collected from 450 human donors across 53 tissues, lymphoblastoid cell lines, and transformed fibroblast cell-lines. We analyzed $16,069$ genes that satisfied filters (e.g. exceeding certain minimum expression levels) that were used during eQTL analyses by the GTEx project (gene list available in `http://stephenslab.github.io/count-clustering/project/utilities/gene_names_all_gtex.txt`).

We fit the GoM model to these data, with number of clusters $K = 5, 10, 15, 20$. For each $K$ we ran the fitting algorithm three times and kept the result with the highest log-likelihood. As might be expected, increasing $K$ highlights finer structure in the data, and for brevity we focus discussion on results for $K = 20$ (Fig 2.1(a)), with results for other $K$ shown in S1. For comparison we also ran several other commonly-used methods for clustering and visualizing gene expression data: Principal Components Analysis (PCA), Multidimensional Scaling (MDS), $t$-Distributed Stochastic Neighbor Embedding ($t$-SNE) [128, 129], and hierarchical clustering (Fig 2.2).

These data present a challenge to visualization and clustering tools, because of both the relatively large number of samples and the complex structure created by the inclusion of many different tissues. Indeed, neither PCA nor MDS provide satisfactory summaries of the structure in these data (Fig 2.2(a,b)): samples from quite different tissues are often superimposed on one another in plots of PC1 vs PC2, and this issue is only partly alleviated by examining more PCs (Supplementary Figure S2). The hierarchical clustering provides perhaps better separation of tissues (Fig 2.2(d)), but producing a clear (static) visualization of the tree is difficult with this many samples. By comparison $t$-SNE (Fig 2.2(b)) and the GoM model (Fig 2.1(a)) both show a much clearer visual separation of samples by tissue, although they achieve this in very different ways. The $t$-SNE representation produces a two-dimensional plot with 20-25 visually-distinct clusters. In contrast, the GoM highlights similarity among samples by assigning them similar membership proportions, resulting in groups of similarly-colored bars in the structure plot. Some tissues are represented by essentially a single cluster/color (e.g. Pancreas, Liver), whereas other tissues are represented as a mixture of multiple clusters (e.g. Thyroid, Spleen). Furthermore, the GoM results highlight biological similarity among some tissues by assigning similar membership proportions to samples from those tissues. For example, samples from several different parts of the brain often have similar memberships, as do the arteries (aorta, tibial and coronary) and skin samples (sun-exposed and un-exposed).

Although it is not surprising that samples cluster by tissue, other results could have occurred. For example, samples could have clustered according to technical variables, such as sequencing batch [44] or sample collection center. While our results do not exclude the possibility that technical variables could have influenced these data, the $t$-SNE and GoM results clearly demonstrate that tissue of origin is the primary source of heterogeneity, and provide a useful initial assurance of data quality.

While in these data both the GoM model and $t$-SNE highlight the primary structure due to tissue of origin, the GoM results have at least two advantages over $t$-SNE. First, the GoM model provides an explicit, quantitative, estimate of the mean expression of each gene in each cluster, making it straightforward to assess which genes and processes drive differences among clusters; see Table 2.1. Reassuringly, many results align with known biology. For example, the purple cluster (cluster 18), which distinguishes Pancreas from other tissues, is enriched for genes responsible for digestion and proteolysis, (e.g. *PRSS1*, *CPA1*, *PNLIP*). Similarly the yellow cluster (cluster 12), which primarily distinguishes Cell EBV Lymphocytes from other tissues, is enriched with genes responsible for immune responses (e.g. *IGHM*, *IGHG1*) and the pink cluster (cluster 19) which mainly appears in Whole Blood, is enriched with genes related hemoglobin complex and oxygen transport (e.g. *HBB*, *HBA1*, *HBA2*). Further, Keratin-related genes characterize the skin cluster (cluster 6, light denim), Myosin-related genes characterize the muscle skeletal cluster (cluster 7, orange), etc. These biological annotations are particularly helpful for understanding instances where a cluster appears in multiple tissues. For example, the top genes in the salmon cluster (cluster 4), which is common to the Gastroesophageal Junction, Esophagus Muscularis and Colon Sigmoid, are related to smooth muscle. And the top genes in the red cluster, highlighted above as common to Breast Mammary tissue, Adipose Subcutaneous and Adipose Visceral, are all related to adipocytes and/or fatty acid synthesis.

A second advantage of the GoM model is that, because it allows partial membership in each cluster, it is better able to highlight partial similarities among distinct tissues. For example, in Figure 2.1(a) the sky blue cluster (cluster 13), appears in testis, pituitary, and thyroid, reflecting shared hormonal-related processes. At the same time, these tissues are distinguished from one another both by their degree of membership in this cluster (testis samples have consistently stronger membership; thyroid samples consistently weaker), and by membership in other clusters. For example, pituitary samples, but not testis or thyroid samples, have membership in the light purple cluster (cluster 2) which is driven by genes related to neurons and synapsis. In the $t$-SNE results these three tissues simply cluster separately into visually distinct groups, with no indication that their expression profiles have something in common (Fig 2.2(b)). Thus, although we find the $t$-SNE results visually attractive, this 2-dimensional projection contains less information than the Structure plot from the GoM (Fig 2.1(a)), which uses color to represent the samples in a 20-dimensional space.

In addition to these qualitative comparisons with other methods, we also used the GTEx data to quantitatively compare the accuracy of the GoM model with hierarchical clustering. Specifically, for each pair of tissues in the GTEx data we assessed whether or not each method correctly partitioned samples into the two tissue groups; see Methods. (Other methods do not provide an explicit clustering of the samples – only a visual representation – and so are not included in these comparisons.) The GoM model was more accurate in this test, succeeding in 88% of comparisons, compared with 79% for hierarchical clustering (Supplemental Figure S3 (c) vs (a)).

### *2.3.2 Sub-analysis of Brain tissues*

Although the analysis of all tissues is useful for assessing global structure, it may miss finer-scale structure within tissues or among similar tissues. For example, here the GoM model applied to all tissues effectively allocated only three clusters to all brain tissues (clusters 1,2 and 9 in Fig 2.1(a)), and we suspected that additional substructure might be uncovered by analyzing the brain samples separately and using more clusters. Fig 2.1(b) shows the Structure plot for $K = 6$ on only the Brain samples. The results highlight much finer-scale structure compared with the global analysis. Brain Cerebellum and Cerebellar hemisphere are essentially assigned to a separate cluster (lime green), which is enriched with genes related to cell periphery and communication (e.g. *PKD1*, *CBLN3*) as well as genes expressed largely in neuronal cells and playing a role in neuron differentiation (e.g. *CHGB*). The spinal cord samples also show consistently strong membership in a single cluster (yellow-orange), the top defining gene for the cluster being *MBP* which is involved in myelination of nerves in the nervous system[56]. Another driving gene, *GFAP*, participates in system development by acting as a marker to distinguish astrocytes during development [7].

The remaining samples all show membership in multiple clusters. Samples from the putamen, caudate and nucleus accumbens show similar profiles, and are distinguished by strong membership in a cluster (cluster 4, bright red) whose top driving gene is *PPP1R1B*, a target for dopamine. And cortex samples are distinguished from others by stronger membership in a cluster (cluster 2, turquoise in Fig 2.1(b)) whose distinctive genes include *ENC1*, which interacts with actin and contributes to the organisation of the cytoskeleton during the specification of neural fate [51].

In comparison, applying PCA, MDS, hierarchical clustering and $t$-SNE to these brain samples reveals less of this finer-scale structure (Supplementary Figures S4). Both PCA and MDS effectively cluster the samples into two groups – those related to the cerebellum vs everything else. Hierarchical clustering also separates out the cerebellum-related tissues from the others, but again the format seems ill-suited to static visualization of more than one thousand samples. For reasons that we do not understand $t$-SNE performs poorly for these data: many samples are allocated to essentially identical locations, and so overplotting obscures them.

### *2.3.3 Single-cell RNA-seq data*

Recently RNA-sequencing has become viable for single cells [123], and this technology has the promise to revolutionize understanding of intra-cellular variation in expression, and regulation more generally [127]. Although it is traditional to describe and categorize cells in terms of distinct cell-types, the actual architecture of cell heterogeneity may be more complex, and in some cases perhaps better captured by the more "continuous" GoM model. In this section we illustrate the potential for the GoM model to be applied to single cell data.

To be applicable to single-cell RNA-seq data, methods must be able to deal with lower sequencing depth than in bulk RNA experiments: single-cell RNA-seq data typically involve substantially lower effective sequencing depth compared with bulk experiments, due to the relatively small number of molecules available to sequence in a single cell. Therefore, as a first step towards demonstrating its potential for single cell analysis, we checked robustness of the GoM model to sequencing depth. Specifically, we repeated the analyses above after thinning the GTEx data by a factor of 100 and $10,000$ to mimic the lower sequencing depth of a typical single cell experiment. For the thinned GTEx data the Structure plot for $K = 20$ preserves most of the major features of the original analysis on unthinned data (Supplemental Figure S5). For the accuracy comparisons with hierarchical clustering, both methods suffer reduced accuracy in thinned data, but the GoM model remains superior (Supplemental Figure S6). For example, when thinning by a factor of $10,000$, the success rate in separating pairs of tissues is 0.32 for the GoM model vs 0.10 for hierarchical clustering.

Having established its robustness to sequencing depth, we now illustrate the GoM model on two single cell RNA-seq datasets: data on mouse spleen from Jaitin *et al* [57] and data on mouse preimplantation embryos from Deng *et al* [22].

## Mouse Spleen data from Jaitin et al, 2014

Jaitin *et al* sequenced over $4,000$ single cells from mouse spleen. Here we analyze $1,041$ of these cells that were categorized as $CD11c+$ in the *sorting markers* column of their data (`http://compgenomics.weizmann.ac.il/tanay/?page_id=519`), and which had total number of reads mapping to non-ERCC genes greater than 600. Our hope was that applying the GoM model to these data would identify, and perhaps refine, the cluster structure evident in [57] (their Fig 2A and 2B). However, the GoM model yielded rather different results (Fig 2.3), where most cells were assigned to have membership in several clusters. Further, the cluster membership vectors showed systematic differences among amplification batches (which in these data is also strongly correlated with sequencing batch). For example, cells in batch 1 are characterized by strong membership in the orange cluster (cluster 5) while those in batch 4 are characterized by strong membership in both the blue and yellow clusters (2 and 6). Some adjacent batches show similar patterns - for example batches 28 and 29 have a similar visual “palette”, as do batches 32-45. And, more generally, these later batches are collectively more similar to one another than they are to the earlier batches (0-4).

The fact that batch effects are detectable in these data is not particularly surprising: there is a growing recognition of the importance of batch effects in high-throughput data generally [21, 71] and in single cell data specifically [44, 52]. And indeed, both clustering methods and the GoM model can be viewed as dimension reduction methods, and such methods can be helpful in controlling for batch effects [72, 118]. However, why these batch effects are not evident in Fig 2A and 2B of [57] is unclear.

## Mouse preimplantation embryo data from Deng et al, 2014

Deng *et al* collected single-cell expression data of mouse preimplantation embryos from the zygote to blastocyst stage [22], with cells from four different embryos sequenced at each stage. The original analysis [22] focuses on trends of allele-specific expression in early embryonic development. Here we use the GoM model to assess the primary structure in these data without regard to allele-specific effects (i.e. combining counts of the two alleles). Visual inspection of the Principal Components Analysis in [22] suggested perhaps 6-7 clusters, and we focus here on results with $K = 6$.

The results from the GoM model (Fig 2.4) clearly highlight changes in expression profiles that occur through early embryonic development stages, and enrichment analysis of the driving genes in each cluster (Table 2.3) indicate that many of these expression changes reflect important biological processes during embryonic preimplantation development.

In more detail: Initially, at the zygote and early 2-cell stages, the embryos are represented by a single cluster (blue in Fig 2.4) that is enriched with genes responsible for germ cell development (e.g., *Bcl2l10* [138], *Spin1* [35]). Moving through subsequent stages the grades of membership evolve to a mixture of blue and magenta clusters (mid 2-cell), a mixture of magenta and yellow clusters (late 2-cell) and a mixture of yellow and green (4-cell stage). The green cluster then becomes more prominent in the 8-cell and 16-cell stages, before dropping substantially in the early and mid-blastocyst stages. That is, we see a progression in the importance of different clusters through these stages, from the blue cluster, moving through magenta and yellow to green. Examining the genes distinguishing each cluster reveals that this progression (blue-magenta-yellow-green) reflects the changing relative importance of several fundamental biological processes. The magenta cluster is driven by genes responsible for the beginning of transcription of zygotic genes (e.g., *Zscan4c-f* show up in the list of top 100 driving genes : see `https://stephenslab.github.io/count-clustering/project/src/deng_cluster_annotations.html`), which takes place in the late 2-cell stage of early mouse embryonic development [36]. The yellow cluster is enriched for genes responsible for heterochromation *Smarcc1* [108] and chromosome stability *Cenpe* [91]. And the green cluster is enriched for cytoskeletal genes (e.g., *Fbxo15*) and cytoplasm genes (e.g., *Tceb1*, *Hsp90ab1*), all of which are essential for compaction at the 8-cell stage and morula formation at the 16-cell stage.

Finally, during the blastocyst stages two new clusters (purple and orange in Fig 2.4) dominate. The orange cluster is enriched with genes involved in the formation of trophectoderm (TE) (e.g., *Tspan8*, *Krt8*, *Id2* [48]), while the purple cluster is enriched with genes responsible for the formation of inner cell mass (ICM) (e.g., *Pdgfra*, *Pyy* [55]).

For comparison, results for PCA, MDS, $t$-SNE and hierarchical clustering are shown in Supplemental Figure S7. All these methods show some clustering structure by preimplantation stage; however only PCA and MDS seem to capture the developmental trajectory from zygote to blastocyst, exhibiting a "horse-shoe" pattern that is expected when similarities among samples approximately reflect an underlying latent ordering [29, 84]. And none of them provide any direct indication of the ICM vs TE structure in the blastocyst samples.

Although the GoM model results clearly highlight some of the key biological processes underlying embryonic preimplantation development, there are also some expected patterns that do not appear. Specifically, just prior to implantation the embryo consists of three different cell types, the trophectoderm (TE), the primitive endoderm (PE), and the epiblast (EPI) [100], with the PE and EPI being formed from the ICM. Thus one might expect the late blastocyst cells to show a clear division into three distinct groups, and for some of the earlier blastocyst cells to show partial membership in one of these groups as they begin to differentiate towards these cell types. Indeed, the GoM model seems well suited to capture this process in principle. However, this is not the result we obtained in practice. In particular, although the two clusters identified by the GoM model in the blastocyst stages appear to correspond roughly to the TE and ICM, even the late blastocyst cells tend to show a gradient of memberships in both these clusters, rather than a clear division into distinct groups. Our results contrast with those from the single-cell mouse preimplantation data of [48], measured by qPCR, where the late blastocyst cells showed a clear visual division into three groups using PCA (their Figure 1).

To better understand the differences between our results for RNA-seq data from [22] and the qPCR results from [48] we applied the GoM model with $K = 3$ to a small subset of the RNA-seq data: the blastocyst cell data at the 48 genes assayed by [48]. These genes were specifically chosen by them to help elucidate cell-fate decisions during early development of the mouse embryo. Still, the GoM model results (Supplemental Figure S8) do not support a clear division of these data into three distinct groups (and neither do PCA or $t$-SNE; Supplemental Figure S9). Rather, the GoM model highlights one cluster (Green in figure), whose membership proportions essentially reflect expression at the *Actb* gene, and two other clusters (Orange and Purple in figure) whose driving genes correspond to genes identified in [48] as being distinctive to TE and EPI cell types respectively. The *Actb* gene is a "housekeeping gene", used by [48] to normalize their qPCR data, and its prominence in the GoM results likely reflects its very high expression levels relative to other genes. However, excluding *Actb* from the analysis still does not lead to a clear separation into three groups (Supplemental Figure S8). Thus, although there are clear commonalities in the structure of these RNA-seq and qPCR data sets, the structure of the single-cell RNA-seq data from [22] is fundamentally more complex (or, perhaps, muddied), and consequently more difficult to interpret.

In addition to trends across development stages, the GoM results also highlight some embryo-level effects in the early stages (Fig 2.4). Specifically, cells from the same embryo sometimes show greater similarity than cells from different embryos. For example, while all cells from the 16-cell stage have high memberships in the green cluster, cells from two of the embryos at this stage have memberships in both the purple and yellow clusters, while the other two embryos have memberships only in the yellow cluster.

The GoM results also highlight a few single cells as outliers. For example, a cell from a 16-cell embryo is represented by the blue cluster - a cluster that represents cells at the zygote and early 2-cell stage. Also, a cell from an 8-stage embryo has strong membership in the purple cluster - a cluster that represents cells from the blastocyst stage. This illustrates the potential for the GoM model to help in quality control: it would seem prudent to consider excluding these outlier cells from subsequent analyses of these data.

## 2.4 Discussion

Our goal here is to highlight the potential for GoM models to elucidate structure in RNA-seq data from both single cell sequencing and bulk sequencing of pooled cells. We also provide tools to identify which genes are most distinctively expressed in each cluster, to aid interpretation of results. As our applications illustrate, the results can provide a richer summary of the structure in RNA-seq data than existing widely-used visualization methods such as PCA and hierarchical clustering. While it could be argued that the GoM model results sometimes raise more questions than they answer, this is exactly the point of an exploratory analysis tool: to highlight issues for investigation, identify anomalies, and generate hypotheses for future testing.

Our results from different methods also highlight another important point: different methods have different strengths and weaknesses, and can compliment one another as well as competing. For example, $t$-SNE seems to provide a much clearer indication of the cluster structure in the full GTEx data than does PCA, but does a poorer job of capturing the ordering of the developmental samples from mouse pre-implantation embryos. While we believe the GoM model often provides a richer summary of the sample structure, we would expect to use it in addition to $t$-SNE and PCA when performing exploratory analyses. (Indeed the methods can be used in combination: both PCA and $t$-SNE can be used to visualize the results of the GoM model, as an alternative or complement to the Structure plot.)

A key feature of the GoM model is that it allows that each sample has a proportion of membership in each cluster, rather than a discrete cluster structure. Consequently it can provide insights into how well a particular dataset really fits a "discrete cluster" model. For example, consider the results for the data from Jaitin *et al* [57] and Deng *et al* [22]: in both cases most samples are assigned to multiple clusters, although the results are closer to "discrete" for the latter than the former. The GoM model is also better able to represent the situation where there is not really a single clustering of the samples, but where samples may cluster differently at different genes. For example, in the GTEx data, the stomach samples share memberships in common with both the pancreas (purple) and the adrenal gland (light green). This pattern can be seen in the Structure plot (Fig 2.1) but not from other methods like PCA, $t$-SNE or hierarchical clustering (Fig 2.2).

Fitting GoM models can be computationally-intensive for large data sets. For the datasets we considered here the computation time ranged from 12 minutes for the data from [22] ($n = 259; K = 6$), through 33 minutes for the data from [57] ($n = 1,041; K = 7$) to $3,370$ minutes for the GTEx data ($n = 8,555; K = 20$). Computation time can be reduced by fitting the model to only the most highly expressed genes, and we often use this strategy to get quick initial results for a dataset. Because these methods are widely used for clustering very large document datasets there is considerable ongoing interest in computational speed-ups for very large datasets, with "on-line" (sequential) approaches capable of dealing with millions of documents [54] that could be useful in the future for very large RNA-seq datasets.

A thorny issue that arises when fitting clustering models is how to select the number of clusters, $K$. Like many software packages for fitting these models, the `maptpx` package implements a measure of model fit that provides one useful guide. However, it is worth remembering that in practice there is unlikely to be a "true" value of $K$, and results from different values of $K$ may complement one another rather than merely competing with one another. For example, seeing how the fitted model evolves as $K$ increases is one way to capture some notion of hierarchy in the clusters identified [99]. More generally it is often fruitful to analyse data in multiple ways using the same tool: for example our GTEx analyses illustrate how analysis of subsets of the data (in this case the brain samples) can complement analyses of the entire data. Finally, as a practical matter, we note that Structure plots can be difficult to read for large $K$ (e.g. $K = 30$) because of the difficulties of choosing a palette with $K$ distinguishable colors.

The version of the GoM model fitted here is relatively simple, and could certainly be embellished. For example, the model allows the expression of each gene in each cluster to be a free parameter, whereas we might expect expression of most genes to be "similar" across clusters. This is analogous to the idea in population genetics applications that allele frequencies in different populations may be similar to one another [37], or in document clustering applications that most words may not differ appreciably in frequency in different topics. In population genetics applications incorporating this idea into the model, by using a correlated prior distribution on these frequencies, can help improve identification of subtle structure [37] and we would expect the same to happen here for RNA-seq data.

Finally, GoM models can be viewed as one of a larger class of "matrix factorization" approaches to understanding structure in data, which also includes PCA, non-negative matrix factorization (NMF), and sparse factor analysis (SFA); see [33]. This observation raises the question of whether methods like SFA might be useful for the kinds of analyses we performed here. (NMF is so closely related to the GoM model that we do not discuss it further; indeed, the GoM model is a type of NMF, because both grades of membership and expression levels within each cluster are required to be non-negative.) Informally, SFA can be thought of as a generalization of the GoM model that allows samples to have *negative* memberships in some "clusters" (actually, "factors"). This additional flexibility should allow SFA to capture certain patterns more easily than the GoM model. For example, a small subset of genes that are over-expressed in some samples and under-expressed in other samples could be captured by a single sparse factor, with positive loadings in the over-expressed samples and negative loadings in the other samples. However, this additional flexibility also comes at a cost of additional complexity in visualizing the results. For example, Supplementary Figures S9, S10, S11 show results of SFA (the version from [33]) for the GTEx data and the mouse preimplantation data: in our opinion, these do not have the simplicity and immediate visual appeal of the GoM model results. Also, applying SFA to RNA-seq data requires several decisions to be made that can greatly impact the results: what transformation of the data to use; what method to induce sparsity (there are many; e.g. [9, 33, 76, 137]); whether to induce sparsity in loadings, factors, or both; etc. Nonetheless, we certainly view SFA as complementing the GoM model as a promising tool for investigating the structure of RNA-seq data, and as a promising area for further work.

## 2.5 Methods and Materials

### *2.5.1 Model Fitting*

We use the `maptpx` R package [121] to fit the GoM model (2.1,2.2), which is also known as "Latent Dirichlet Allocation" (LDA). The `maptpx` package fits this model using an EM algorithm to perform Maximum a posteriori (MAP) estimation of the parameters $q$ and $\theta$. See [121] for details.

### *2.5.2 Visualizing Results*

In addition to the Structure plot, we have also found it useful to visualize results using t-distributed Stochastic Neighbor Embedding (t-SNE), which is a method for visualizing high dimensional datasets by placing them in a two dimensional space, attempting to preserve the relative distance between nearby samples [128, 129]. Compared with the Structure plot our t-SNE plots contain less information, but can better emphasize clustering of samples that have similar membership proportions in many clusters. Specifically, t-SNE tends to place samples with similar membership proportions together in the two-dimensional plot, forming visual "clusters" that can be identified by eye (e.g. `http://stephenslab.github.io/count-clustering/project/src/tissues_tSNE_2.html`). This may be particularly helpful in settings where no external information is available to aid in making an informative Structure plot.

### *2.5.3 Cluster annotation*

To help biologically interpret the clusters, we developed a method to identify which genes are most distinctively differentially expressed in each cluster. (This is analogous to identifying "ancestry informative markers" in population genetics applications [98].) Specifically, for each cluster $k$ we measure the distinctiveness of gene $g$ with respect to any other cluster $l$ using

$$KL^g[k,l] := \theta_{kg} \, log \frac{\theta_{kg}}{\theta_{lg}} + \theta_{lg} - \theta_{kg}, \tag{2.3}$$

which is the Kullback–Leibler divergence of the Poisson distribution with parameter $\theta_{kg}$ to the Poisson distribution with parameter $\theta_{lg}$. For each cluster $k$, we then define the distinctiveness of gene $g$ as

$$D^g[k] = \min_{l \neq k} KL^g[k,l]. \tag{2.4}$$

The higher $D^g[k]$, the larger the role of gene $g$ in distinguishing cluster $k$ from all other clusters. Thus, for each cluster $k$ we identify the genes with highest $D^g[k]$ as the genes driving the cluster $k$. We annotate the biological functions of these individual genes using the `mygene` R Bioconductor package [124].

For each cluster $k$, we filter out a number of genes (top 100 for the Deng *et al* data [22] and GTEx V6 data [19]) with highest $D^g[k]$ value and perform a gene set over-representation analysis of these genes against all the other genes in the data representing the background. To do this, we used ConsensusPathDB database (`http://cpdb.molgen.mpg.de/`) [62] [88]. See Table 2.1-2.2 and Table 2.3 for the top significant gene ontologies driving each cluster in the GTEx V6 data and the Deng *et al* data respectively.

### *2.5.4 Comparison with hierarchical clustering*

We compared the GoM model with a distance-based hierarchical clustering algorithm by applying both methods to samples from pairs of tissues from the GTEx project, and assessed their accuracy in separating samples according to tissue. For each pair of tissues we randomly selected 50 samples from the pool of all samples coming from these tissues. For the hierarchical clustering approach we cut the dendrogram at $K = 2$, and checked whether or not this cut partitions the samples into the two tissue groups. (We applied hierarchical clustering using Euclidean distance, with both complete and average linkage; results were similar and so we showed results only for complete linkage.)

For the GoM model we analysed the data with $K = 2$, and sorted the samples by their membership in cluster 1. We then partitioned the samples at the point of the steepest fall in this membership, and again we checked whether this cut partitions the samples into the two tissue groups. Supplemental Figure S3 shows, for each pair of tissues, whether each method successfully partitioned the samples into the two tissue groups.

### *2.5.5 Thinning*

We used "thinning" to simulate lower-coverage data from the original higher-coverage data.. Specifically, if $c_{ng}$ is the counts of number of reads mapping to gene $g$ for sample $n$ for the original data, we simulated thinned counts $t_{ng}$ using

$$t_{ng} \sim Bin(c_{ng}, p_{thin}) \tag{2.5}$$

where $p_{thin}$ is a specified thinning parameter.

### *2.5.6 Code Availability*

Our methods are implemented in an R package `CountClust`, available as part of the Bioconductor project at `https://www.bioconductor.org/packages/3.3/bioc/html/CountClust.html`. The development version of the package is also available at `https://github.com/kkdey/CountClust`. Code for reproducing results reported here is available at `http://`

`stephenslab.github.io/count-clustering/`.

## 2.6 Author contributions

Dey, KK and Stephens, M designed the method. Dey, KK implemented the method. Dey, KK and Hsiao, CJ ran the experiments. Dey, KK and Hsiao, CJ produced the figures. Dey, KK, Hsiao, CJ and Stephens, M wrote the paper.

Table 2.1: **Cluster Annotations GTEx V6 data (with GO annotations).**

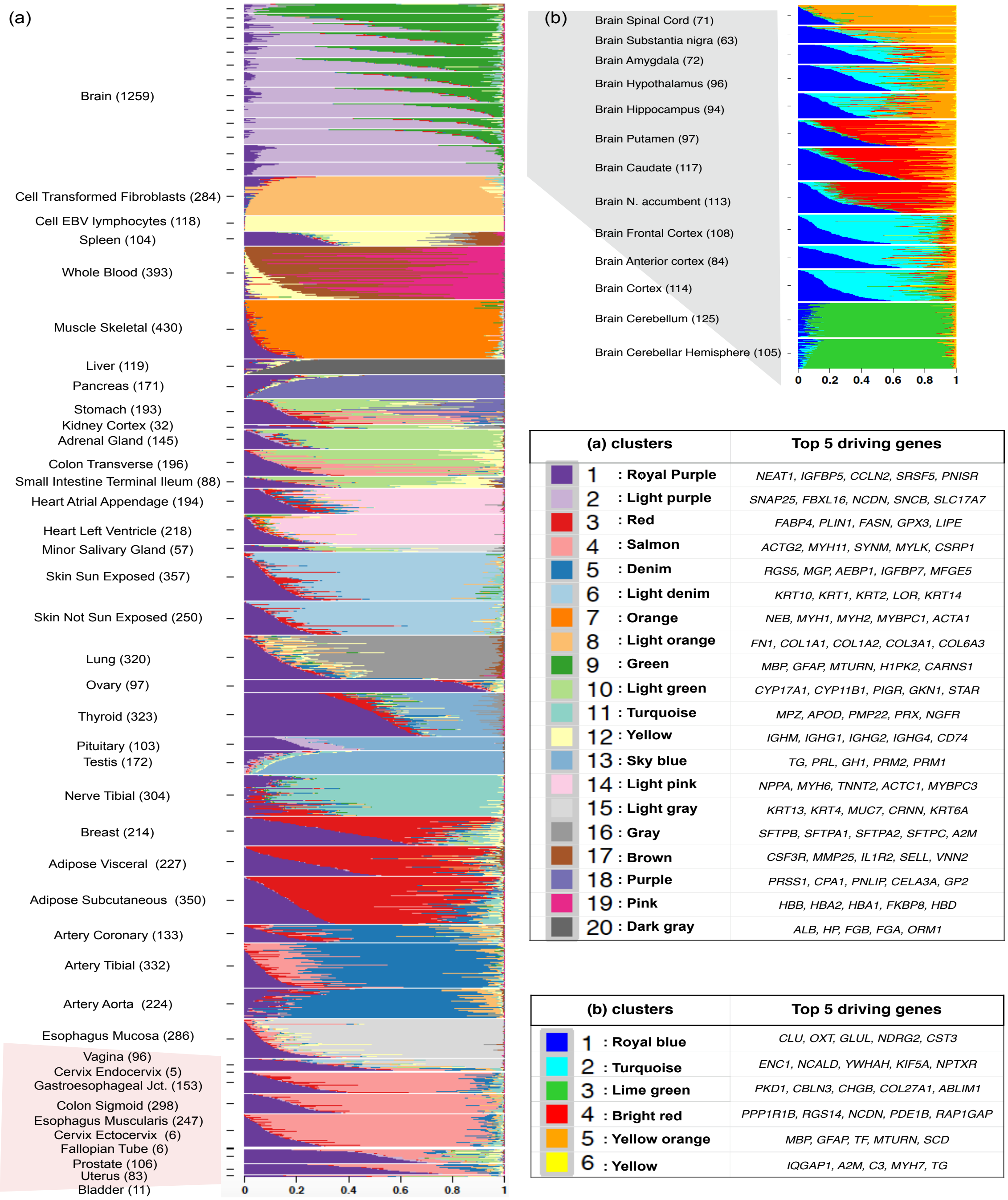
| Cluster | Top 5 Driving Genes | Top significant GO terms (function)[q-value] |
|---|---|---|
| 1. Royal purple | *NEAT1, IGFBP5, CCLN2, SRSF5, PNISR* | GO:0005654 (nucleoplasm)[2e-10], GO:0044822 (poly-A RNA binding)[3e-09], GO:0044428 (nuclear part)[1e-09], GO:0043233 (organelle lumen)[2e-08] |
| 2. Light purple | *SNAP25, FBXL16, NCDN, SNCB, SLC17A7* | GO:0097458 (neuron part)[2e-25], GO:0007268 (synaptic transmission)[9e-18], GO:0030182 (neuron differentiation)[2e-14], GO:0022008 (neurogenesis)[1e-13], GO:0007267 (cell-cell signaling)[3e-13] |
| 3. Red | *FABP4, PLIN1, FASN, GPX3, LIPE* | GO:0044255 (cellular lipid metabolism)[1e-09], GO:0006629 (lipid metabolism)[1e-09], GO:0006639 (acylglycerol metabolism)[3e-08], GO:0045765 (angiogenesis regulation)[4e-08] |
| 4. Salmon | *ACTG2, MYH11, SYNM, MYLK, CSRP1* | GO:0043292 (contractile fiber)[3e-13], GO:0006936 (muscle contraction)[5e-12], GO:0030016 (myofibril)[5e-12], GO:0015629 (actin cytoskeleton)[2e-12], GO:0005925 (focal adhesion)[6e-11] |
| 5. Denim | *RGS5, MGP, AEBP1, IGFBP7, MFGE8* | GO:0005578 (proteinaceous extracellular matrix)[4e-20], GO:0030198 (extracellular matrix)[2e-18], GO:0007155 (cell adhesion)[4e-14], GO:0001568 (blood vessel development)[4e-13] |
| 6. Light denim | *KRT10, KRT1, KRT2, LOR, KRT14* | GO:0008544 (epidermis development)[3e-12], GO:0043588 (skin development)[5e-10], GO:0042303 (molting cycle)[8e-06], GO:0042633 (hair cycle)[7e-06], GO:0048513 (organ development)[6e-05] |
| 7. Orange | *NEB, MYH1, MYH2, MYBPC1, ACTA1* | GO:0043292 (contractile fiber)[6e-52], GO:0030016 (myofibril)[1e-51], GO:0030017 (sarcomere)[5e-40], GO:0003012 (muscle system process)[2e-25], GO:0015629 (actin cytoskeleton)[1e-25] |
| 8. Light orange | *FN1, COL1A1, COL1A2, COL3A1, COL6A3* | GO:0030198 (extracellular matrix)[6e-29], GO:0043062 (extracellular structure)[4e-29], GO:0032963 (collagen metabolism)[3e-16], GO:0030199 (collagen fibril organization)[1e-14], GO:0030574 (collagen catabolism)[1e-14] |
| 9. Green | *MBP, GFAP, MTURN, HIPK2, CARNS1* | GO:0043209 (myelin sheath)[4e-07], GO:0007399 (nervous system development)[4e-05], GO:0008366 (axon ensheathment)[9e-05], GO:0044430 (cytoskeletal part)[1e-04], GO:0005874 (microtubule)[3e-04] |
| 10. Light green | *CYP17A1, CYP11B1, PIGR, GKN1, STAR* | GO:0006694 (steroid biosynthesis)[2e-13], GO:0008202 (steroid metabolism)[1e-12], GO:0016125 (sterol metabolism)[1e-11], GO:0042446 (hormone biosynthesis)[1e-10], GO:0008207 (C21-steroid hormone metabolism)[3e-10] |
| 11. Turquoise | *MPZ, APOD, PMP22, PRX, NGFR* | GO:0007272 (ensheathment of neurons)[4e-07], GO:0008366 (axon ensheathment)[7e-07], GO:0042552 (myelination)[7e-06], GO:0048856 (anatomical structure development)[1e-06], GO:0005578 (proteinaceous extracellular matrix)[1e-06] |
| 12. Yellow | *IGHM, IGHG1, IGHG2, IGHG4, CD74* | GO:0006955 (immune response)[1e-18], GO:0002252 (immune effector process)[7e-18], GO:0003823 (antigen binding)[1e-15], GO:0019724 (B-cell mediated immunity)[5e-13], GO:0002684 (positive regulation immune system)[6e-13] |
| 13. Sky blue | *TG, PRL, GH1, PRM2, PRM1* | GO:0019953 (sexual reproduction)[8e-10], GO:0048232 (male gamete generation)[2e-08], GO:0035686 (sperm fibrous sheath)[4e-06], GO:0005179 (hormone activity)[6e-05], GO:0042403 (thyroid hormone metabolism)[2e-04] |
| 14. Light pink | *NPPA, MYH6, TNNT2, ACTC1, MYBPC3* | GO:0045333 (cellular respiration)[2e-34], GO:0022904 (respiratory electron transport)[8e-33], GO:0015980 (energy derivation by oxidation of organic compounds)[4e-30], GO:0031966 (mitochondrial membrane)[5e-26] |
| 15. Light gray | *KRT13, KRT4, MUC7, CRNN, KRT6A* | GO:0070062 (extracellular exosome)[2e-23], GO:0043230 (extracellular organelle)[3e-23], GO:0031982 (vesicle)[3e-20], GO:0008544 (epidermis development)[2e-18], GO:0043588 (skin development)[1e-13] |
| 16. Gray | *SFTPBβ, SFTPA1, SFTPA2, SFTPC, A2M* | GO:0001525 (angiogenesis)[5e-08], GO:0001944 (vasculature development)[2e-07], GO:0048514 (blood vessel morphogenesis)[2e-07], GO:0040012 (locomotion regulation)[4e-06], GO:2000145 (cell motility)[1e-05] |
| 17. Brown | *CSF3R, MMP25, IL1R2, SELL, VNN2* | GO:0006955 (immune response)[8e-22], GO:0006952 (defense response)[9e-16], GO:0071944 (cell periphery)[7e-15], GO:0005886 (plasma membrane)[7e-15], GO:0050776 (regulation of immune response)[2e-12] |
| 18. Purple | *PRSS1, CPA1, PNLIP, CELA3A, GP2* | GO:0007586 (digestion)[3e-14], GO:0004252 (serine-type endopeptidase activity)[4e-08], GO:0006508 (proteolysis)[6e-06], GO:0016787 (hydrolase activity)[6e-05], GO:0044241 (lipid digestion)[1e-04] |
| 19. Pink | *HBB, HBA2, HBA1, FKBP8, HBD* | GO:0005833 (hemoglobin complex)[1e-13], GO:0015669 (gas transport)[4e-11], GO:0020037 (heme binding)[3e-07], GO:0031720 (haptoglobin binding)[3e-06], GO:0006950 (response to stress)[6e-04] |
| 20. Dark gray | *ALB, HP, FGB, FGA, ORM1* | GO:0072562 (blood microparticle)[1e-27], GO:0043230 (extracellular organelle)[1e-24], GO:0044710 (single organism metabolism)[7e-20], GO:0019752 (carboxylic acid metabolism)[1e-18], GO:0034364 (high density lipoprotein)[3e-16] |

Table 2.2: **Cluster Annotations for GTEx V6 Brain data.**

| Cluster | Top 5 Driving Genes | Top significant GO terms |
|---|---|---|
| 1. Royal blue | *CLU, OXT, GLUL, NDRG2, CST3* | GO:0043230 (extracellular organelle)[5e-11], GO:1903561 (extracellular vesicle)[6e-11], GO:0070062 (extracellular exosome)[2e-09], GO:0006950 (response to stress)[3e-10], GO:0031988 (membrane bound vesicle)[1e-10] |
| 2. Turquoise | *ENC1, NCALD, YWHAH, KIF5A, NPTXR* | GO:0097458 (neuron part)[3e-11], GO:0008092 (cytoskeletal protein binding)[7e-11], GO:0031175 (neuron projection development)[7e-09], GO:0030182 (neuron differentiation)[4e-08], GO:0007268 (synaptic transmission)[1e-08] |
| 3. Lime green | *PKD1, CBLN3, CHGB, COL27A1, ABLIM1* | GO:0005089 (Rho guanyl-nucleotide exchange factor activity)[1e-03], GO:0016604 (nuclear body)[0.002], GO:0022008 (neurogenesis)[0.02], GO:0035239 (tube morphogenesis)[0.08], GO:0007269 (neurotransmitter secretion)[0.10] |
| 4. Red | *PPP1R1B, RGS14, NCDN, PDE1B, RAP1GAP* | GO:0065009 (regulation of molecular function)[2e-06], GO:0036477 (somatodendritic compartment)[6e-05], GO:0007268 (synaptic transmission)[1e-03], GO:0023051 (signaling regulation)[2e-03], GO:0010646 (cell communication regulation)[1e-03] |
| 5. Yellow orange | *MBP, GFAP, TF, MTURN, SCD* | GO:0043209 (myelin sheath)[2e-09], GO:0007399 (nervous system development)[1e-04], GO:0005737 (cytoplasm)[1e-04], GO:0048471 (perinuclear region of cytoplasm)[5e-04], GO:0007272 (ensheathment of neurons)[1e-02] |
| 6. Yellow | *IQGAP1, A2M, C3, MYH7, TG* | GO:0072562 (blood microparticle)[1e-10], GO:0044449 (contractile fiber part)[1e-10], GO:0043230 (extracellular organelle)[7e-10], GO:0030017 (sarcomere)[1e-08], GO:0072376 (protein activation cascade)[1e-08] |

Table 2.3: **Cluster Annotations for Deng et al (2014) data.**

| Cluster | Top 10 Driving Genes | Top significant GO terms |
|---|---|---|
| 1. Blue | *Bcl2l10*, *E330034G19Rik*, *Tcl1*,*LOC100502936*, *Oas1d*, *AU022751*, *Spin1*, *Khdc1b*, *D6Ertd527e*, *Btg4* | GO:0007276 (gamete generation)[7e-06], GO:0032504 (multicellular organism reproduction)[3e-06], GO:0044702 (single organism reproduction)[2e-05], GO:0048477 (oogenesis)[5e-04], GO:0048599 (oocyte development)[1e-03], GO:0009994 (oocyte differentiation)[1e-03] |
| 2. Magenta | *Obox3*, *Zfp352*, *Gm8300*, *Usp17l5*, *BB287469*, *Rfpl4b*, *Gm2022*, *Gm5662*, *Gm11544* , *Gm4850* | GO:0016604 (nuclear body)[1e-04], GO:0005814 (centriole)[4e-03], GO:0044450 (microtubule organizing center part) [8e-03] |
| 3. Yellow | *Rtn2*, *Ebna1bp2*, *Zfp259*, *Nasp*, *Cenpe*, *Rnf216*, *Ctsl*, *Tor1b*, *Ankrd10*, *Lamp2* | GO:0044428 (nuclear part)[1e-08], GO:0031981 (nuclear lumen)[3e-08], GO:0070013 (intracellular organelle lumen)[9e-08], GO:0005730 (nucleolus)[5e-07], GO:0005654 (nucleoplasm)[4e-05], GO:0003723 (RNA binding)[1e-04] |
| 4. Green | *Timd2*, *Isyna1*, *Alppl2*, *Prame15*, *Fbxo15*, *Tceb1*, *Gpd1l*, *Pemt*, *Hsp90aa1*, *Hsp90ab1* | GO:0005829 (cytosol)[4e-10], GO:0044444 (cytoplasmic part)[2e-05], GO:1901575 (organic substance catabolic process)[9e-04], GO:0000151 (ubiquitin ligase com- plex)[1e-04], GO:0009056 (catabolic process)[1e-03], GO:0044265 (cellular macromolecule catabolic process)[1e-03], GO:0051082 (unfolded protein binding)[9e-04] |
| 5. Purple | *Upp1*, *Tdgf1*, *Aqp8*, *Fabp5*, *Tat*, *Pdgfra*, *Pyy*, *Prdx1*, *Col4a1*, *Spp1* | GO:0044710 (single-organism metabolic process) [1e-05], GO:0006950 (response to stress) [1e-05], GO:0070062 (extracellular exosome)[1e-05], GO:0043230 (extracellular organelle)[2e-05], GO:1903561 (extracellular vesicle)[1e-05], GO:0006979 (response to oxidative stress)[7e-04], GO:0048514 (blood vessel morphogenesis)[7e-04], GO:0001944 (vasculature development)[3e-03] |
| 6. Orange | *Actb*, *Krt18*, *Fabp3*, *Id2*, *Tspan8*, *Gm2a*, *Lgals1*, *Adh1* , *Lrp2*, *BC051665* | GO:0065010 (extracellular membrane-bounded organelle), GO:0070062 (extracellular exosome)[4e-23], GO:0043230 (extracellular organelle)[5e-23], GO:1903561 (extracellular vesicle)[3e-23], GO:0031982 (vesicle)[4e-18], GO:0030036 (actin cytoskeleton and organization)[4e-12], GO:0032432 (actin filament bundle)[2e-09], GO:0005912 (adherens junction)[2e-09] |



| (a) clusters | Top 5 driving genes |
|---|---|
| 🟪 1 : Royal Purple | *NEAT1, IGFBP5, CCLN2, SRSF5, PNISR* |
| 🟪 2 : Light purple | *SNAP25, FBXL16, NCDN, SNCB, SLC17A7* |
| 🟥 3 : Red | *FABP4, PLIN1, FASN, GPX3, LIPE* |
| 🟥 4 : Salmon | *ACTG2, MYH11, SYNM, MYLK, CSRP1* |
| 🟦 5 : Denim | *RGS5, MGP, AEBP1, IGFBP7, MFGE5* |
| 🟦 6 : Light denim | *KRT10, KRT1, KRT2, LOR, KRT14* |
| 🟧 7 : Orange | *NEB, MYH1, MYH2, MYBPC1, ACTA1* |
| 🟧 8 : Light orange | *FN1, COL1A1, COL1A2, COL3A1, COL6A3* |
| 🟩 9 : Green | *MBP, GFAP, MTURN, H1PK2, CARNS1* |
| 🟩 10 : Light green | *CYP17A1, CYP11B1, PIGR, GKN1, STAR* |
| 🟩 11 : Turquoise | *MPZ, APOD, PMP22, PRX, NGFR* |
| 🟨 12 : Yellow | *IGHM, IGHG1, IGHG2, IGHG4, CD74* |
| 🟦 13 : Sky blue | *TG, PRL, GH1, PRM2, PRM1* |
| 🟪 14 : Light pink | *NPPA, MYH6, TNNT2, ACTC1, MYBPC3* |
| ⬜ 15 : Light gray | *KRT13, KRT4, MUC7, CRNN, KRT6A* |
| ⬛ 16 : Gray | *SFTPB, SFTPA1, SFTPA2, SFTPC, A2M* |
| 🟫 17 : Brown | *CSF3R, MMP25, IL1R2, SELL, VNN2* |
| 🟪 18 : Purple | *PRSS1, CPA1, PNLIP, CELA3A, GP2* |
| 🟥 19 : Pink | *HBB, HBA2, HBA1, FKBP8, HBD* |
| ⬛ 20 : Dark gray | *ALB, HP, FGB, FGA, ORM1* |

| (b) clusters | Top 5 driving genes |
|---|---|
| 🟦 1 : Royal blue | *CLU, OXT, GLUL, NDRG2, CST3* |
| 🟦 2 : Turquoise | *ENC1, NCALD, YWHAH, KIF5A, NPTXR* |
| 🟩 3 : Lime green | *PKD1, CBLN3, CHGB, COL27A1, ABLIM1* |
| 🟥 4 : Bright red | *PPP1R1B, RGS14, NCDN, PDE1B, RAP1GAP* |
| 🟧 5 : Yellow orange | *MBP, GFAP, TF, MTURN, SCD* |
| 🟨 6 : Yellow | *IQGAP1, A2M, C3, MYH7, TG* |

Figure 2.1: **(a)**: Structure plot of estimated membership proportions for GoM model with $K = 20$ clusters fit to 8555 tissue samples from 53 tissues in GTEx data. Each horizontal bar shows the cluster membership proportions for a single sample, ordered so that samples from the same tissue are adjacent to one another. **(b)**: Structure plot of estimated membership proportions for $K = 4$ clusters fit to only the brain tissue samples.

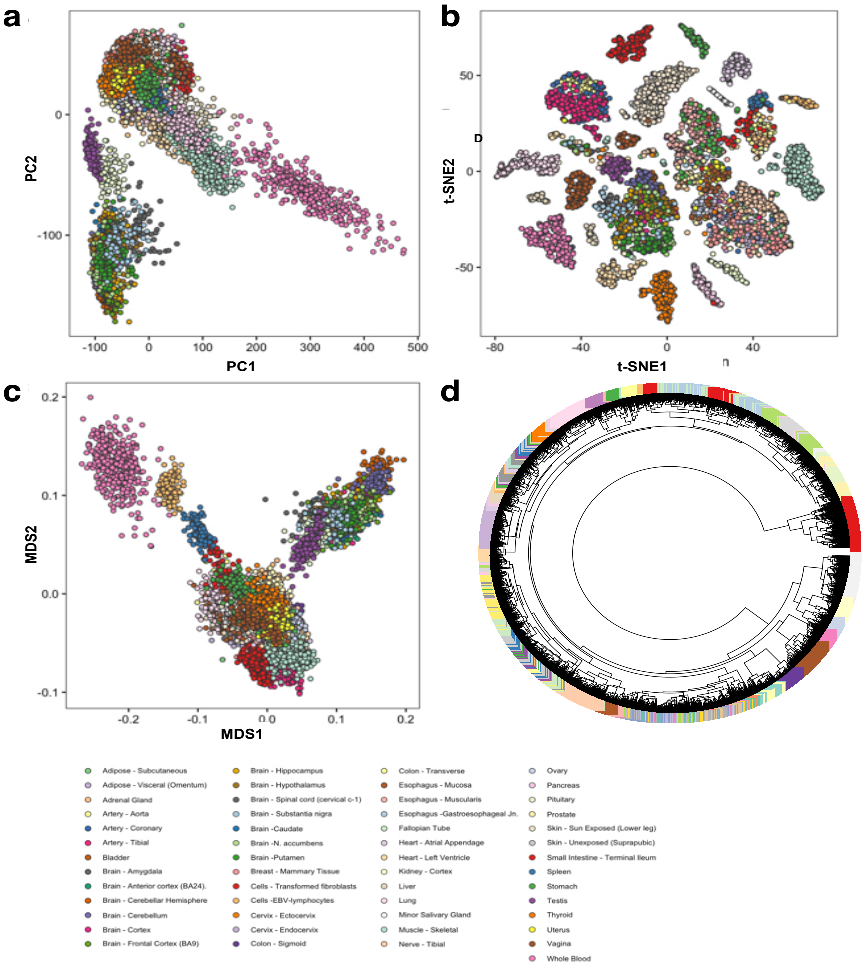


Figure 2.2: Visualization of the same GTEx data as in Figure 1 (a) across all tissues using standard and widely used approaches - Principal Component Analysis (PCA), Multi dimensional Scaling (MDS), t-SNE and hierarchical clustering. All the analysis are done on log CPM normalized expression data to remove library size effects. **(a)**: Plot of PC1 vs PC2 on the log CPM expression data, **(b)**: Plot of first two dimensions of the t-SNE plot, **(c)** Plot of first two dimensions of the Multi-Dimensional Scaling (MDS) plot. **(d)** Dendrogram for the hierarchical clustring of the GTEx tissue samples based on the log CPM expression data.

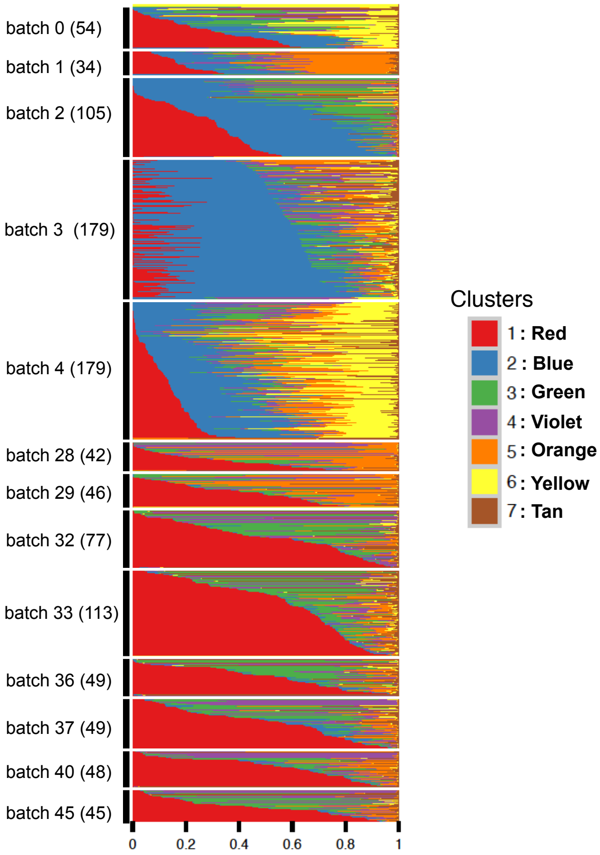
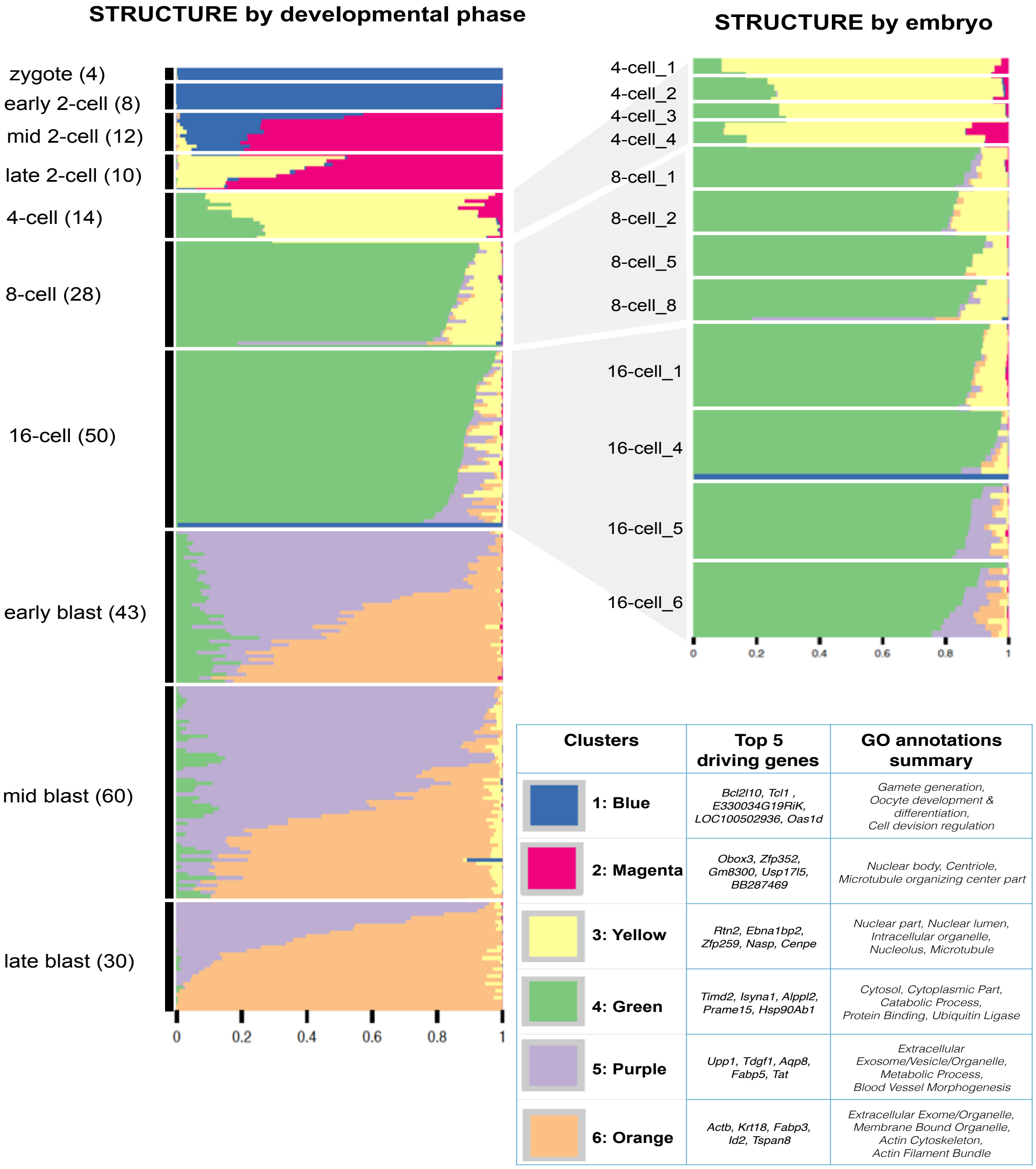


Figure 2.3: Structure plot of estimated membership proportions for GoM model with $K = 7$ clusters fit to $1,041$ single cells from [57]. The samples (cells) are ordered so that samples from the same amplification batch are adjacent and within each batch, the samples are sorted by the proportional representation of the underlying clusters. In this analysis the samples do not appear to form clearly-defined clusters, with each sample being allocated membership in several "clusters". Membership proportions are correlated with batch, and some groups of batches (e.g. 28-29; 32-45) show similar palettes. These results suggest that batch effects are likely influencing the inferred structure in these data.



| Clusters | Top 5 driving genes | GO annotations summary |
|---|---|---|
| 🟦 1: Blue | *Bcl2l10, Tcl1 , E330034G19RiK, LOC100502936, Oas1d* | *Gamete generation, Oocyte development & differentiation, Cell devision regulation* |
| 🟥 2: Magenta | *Obox3, Zfp352, Gm8300, Usp17l5, BB287469* | *Nuclear body, Centriole, Microtubule organizing center part* |
| 🟨 3: Yellow | *Rtn2, Ebna1bp2, Zfp259, Nasp, Cenpe* | *Nuclear part, Nuclear lumen, Intracellular organelle, Nucleolus, Microtubule* |
| 🟩 4: Green | *Timd2, Isyna1, Alppl2, Prame15, Hsp90Ab1* | *Cytosol, Cytoplasmic Part, Catabolic Process, Protein Binding, Ubiquitin Ligase* |
| 🟪 5: Purple | *Upp1, Tdgf1, Aqp8, Fabp5, Tat* | *Extracellular Exosome/Vesicle/Organelle, Metabolic Process, Blood Vessel Morphogenesis* |
| 🟧 6: Orange | *Actb, Krt18, Fabp3, Id2, Tspan8* | *Extracellular Exome/Organelle, Membrane Bound Organelle, Actin Cytoskeleton, Actin Filament Bundle* |

Figure 2.4: Structure plot of estimated membership proportions for GoM model with $K = 6$ clusters fit to 259 single cells from [22]. Each cluster is annotated by the genes that are most distinctively expressed in that cluster, and by the gene ontology categories for which these distinctive genes are most enriched.

# Chapter 3

# *aRchaic* : Modeling and Visualization of DNA damage patterns using a Grade of Membership Model

*(with H. Al-Asadi, J. Novembre and M. Stephens)*

## 3.1 Introduction

Ancient DNA (aDNA) research has seen rapid growth with the recent advancements in recovery of short DNA fragments, increased throughput, and lower per-base cost in sequencing [112]. Both the number and size of aDNA datasets have grown rapidly over the last five years, and several recent studies sequenced hundreds of ancient individuals [6, 78, 79, 85].

This rapid recent growth in aDNA research has provided many new insights into human history. However, working with aDNA remains challenging. For example, ancient samples often contain very little endogenous DNA because in many environments DNA degrades rapidly post-mortem [104]. Furthermore, ancient samples are often contaminated by microbes and exogenous human DNA [77]. Both these factors mean that many sequence reads generated by an aDNA study may not actually come from the ancient sample.

Because of these challenges aDNA researchers pay careful attention to quality control (QC), including checking sequencing reads from each sample for signatures of endogenous aDNA. These signatures include: short fragment length, an enrichment of purines before strand breaks, and a high frequency of cytosine to thymine substitutions ($C \rightarrow T$) at the ends of fragments [16, 45, 60, 104, 116]. One common QC procedure is to plot, for each sample, the $C \rightarrow T$ mismatch rate as a function of position from the end of the read, and to look for an elevated rate near the ends of reads as an indication of the presence of endogenous aDNA. Another common procedure is to look for elevated rates of purines (A & G) before the 5' strand-breaks. Both these procedures are implemented in the software *mapdamage* [45, 60], for example.

These commonly-used QC checks, though simple and useful, have several limitations. For example, they produce a plot for each sample, which can be inconvenient to work with and difficult to compare across many samples. This issue becomes increasingly important with the growing size of aDNA datasets. These plots can also be difficult to interpret, in part because they do not contrast observed patterns with expected patterns in modern samples. Finally, these approaches can detect only pre-defined damage signatures, and may fail to capture other key features or artifacts in the data.

Here we introduce methods to help address these problems. These methods start with a Binary Alignment Map (BAM) file, obtained by aligning each read to a reference genome. The BAM file includes information on the mismatches that occur in each read (vs the reference). We characterize each mismatch by several relevant features, including its type (e.g. $C \rightarrow G$, etc), flanking bases, and distance from the end of the read. We then use these features to cluster the mismatches into groups, which we call *mismatch profiles.* Intuitively, a mismatch profile associated with post-mortem damage is expected to show high levels of $C \rightarrow T$ mismatches at the ends of reads. On the other hand, a mismatch profile that is typical of modern DNA polymorphism will show a different pattern, such as a transition to transversion ratio of 2:1 [46]. Finally we estimate the relative frequency of each mismatch profile in each sample, which we refer to as the "Grade of membership" [34] of that sample in that mismatch profile. These grades of membership should reflect which processes generated mismatches in each sample. For example ancient samples should have a high grade of membership in mismatch profiles characteristic of post-mortem DNA damage. Grade of membership models are widely used to infer structure in admixed populations [89], document collections [15], RNA-seq data [24] and somatic mutation data [115] for example.

We have implemented methods to fit this model, and visualize the results in a software package, aRchaic. For example, the grades of membership for all samples are succinctly displayed in a single STRUCTURE plot [99], and each mismatch profile is displayed using simple intuitive plots [26]. Together these plots provide a concise visual summary of DNA damage patterns, as well as other processes generating mismatches in the data.

## 3.2 Methods

For each sample $i = 1, \ldots, I$ we first obtain a BAM file. From this BAM file we extract information on the mismatches (vs a reference) that occur in reads from the sample. First we filter out low-quality reads (mapping $\leq 30$), low-quality mismatches (base quality $\leq 20$), and mismatches that occur more than 20bp from the end of a read (since these are unlikely to reflect damage patterns []). When a read carries more than one mismatch we treat these as independent (an assumption we verified by checking that the probability of a mismatch conditional on the occurrence of another mismatch on the same read is not significantly different from the marginal probability, p-value = 0.43).

Let $J_i$ denote the total number of remaining mismatches, and for each mismatch $j = 1, \cdots, J_i$ let $x_{i,j} = (x_{i,j,1}, x_{i,j,2}, x_{i,j,3}, x_{i,j,4}, x_{i,j,5})$ denote the following features:

- $x_{i,j,1} \in \{T \to A,\ T \to G,\ T \to C,\ C \to T,\ C \to A,\ C \to G\}$ denotes the mismatch.
- $x_{i,j,2} \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ denotes the nucleotides immediately 5' to the mismatch on the reference genome.
- $x_{i,j,3} \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ denotes the nucleotides immediately 3' to the mismatch on the reference genome.
- $x_{i,j,4} \in \{1, ...20\}$ denotes the distance from the mismatch to the (nearest) end of the read.

- $x_{i,j,5} \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ denotes the nucleotide of the base directly 5' to the position of the strand break of the read.

These features are designed to reflect the major modes of DNA damage ([16, 90, 104, 104]). For each feature $l \in \{1, \ldots, 5\}$ we let $M_l$ denote the number of possible values of $x_{i,j,l}$, and for notational convenience we treat $x_{i,j,l}$ as being in $\{1, \ldots, M_l\}$. For example we represent $x_{i,j,1} = T \rightarrow A$ by $x_{i,j,1} = 1$.

Our model assumes that each mismatch in each individual arose from one of $K$ mismatch profiles ("clusters"). We introduce latent variables $z_{i,j} \in \{1, .., K\}$ to denote the profile that gave rise to mismatch $j$ in individual $i$. We assume

$$\Pr(z_{i,j} = k) = q_{i,k}, \tag{3.1}$$

where $q_{i,k}$ represents the membership proportion of individual $i$ in mismatch profile (cluster) $k \in 1, .., K$.

We further assume that, given $z_{i,j} = k$,

$$\Pr(x_{i,j,l} = m | z_{i,j} = k) = f_{k,l}(m), \tag{3.2}$$

where $m \in \{1, \ldots, M_l\}$, and $f_{k,l}(m)$ denotes the relative frequency of $m$ at feature $l$ in cluster $k$. We follow [115] in assuming independence among features within each cluster.

Putting this all together, and assuming independence of observations yields the likelihood:

$$L(q, f; x) = \prod_{i,j,l} \sum_{k} f_{k,l}(x_{i,j,l}) q_{i,k}. \tag{3.3}$$

We fit this model, and estimate the individual parameters ($q$) and cluster parameters ($f$) by maximum likelihood using an accelerated EM algorithm. We use the same EM updates as in equations 2-4 in [115], and we add first-order quasi-Newton acceleration to improve convergence [3, 67, 121].

For each cluster $k$, we visualize the cluster parameters $f_k$ using as an *EDLogo* plot [26]. The *EDLogo* plot allows one to visualize both enrichment and depletion of mismatch features scaled against a reference frequency. In our application, the reference frequency was computed from individuals in the 1000 Genomes Project because we wanted to compare mismatch profiles in our samples against that of modern individuals [1]. We use a STRUCTURE plot [99] to visualize the estimates of $q_{i,k}$ for each sample.

## 3.3 Results

We demonstrate the utility of `aRchaic` using three case-studies.

### *3.3.1 `aRchaic` clustering of modern and ancient individuals*

We applied `aRchaic` to a combined dataset of xx ancient samples from four recent studies [42, 68, 79, 116] and 60 modern samples from the 1000 Genomes Project [1] (n=50) and HGDP [] (n=10). Two of the aDNA studies used partial-UDG treated libraries, which removes most – but not all – of the damage [97].

Figure 3.2 shows results from `aRchaic` with $K = 3$ (see Supplementary Figure S12 for $K = 4, 5, 6$). To give a sense of computational requirements, these results took approximately 23 minutes on a single modern compute node. The results clearly highlight differences between modern, ancient (UDG), and ancient (non-UDG) samples. The modern samples show very strong membership in a single cluster (red). As expected, this ”modern” cluster shows only modest enrichment in its mismatch type, flanking base composition, and mismatch location, relative to the modern background.

Ancient (non-UDG) samples show high membership in a second cluster (blue). This cluster is characterized by a very strong enrichment of $C \to T$ mismatches at the ends of the reads, which is a typical sign of DNA damage [97]. This enrichment is also accompanied by a depletion of guanine just 3' to the mismatch, which likely reflect the fact that the CpG combination occurs less often than expected by chance [113].

The ancient UDG-treated individuals show high membership in both the red cluster and a third (orange) cluster. The high membership in the red cluster presumably reflects the fact that the UDG-treatment repairs much of the damage in these samples, making them look more "modern" in their mismatch profiles. The orange cluster is characterized by enrichment of $C \to T$ mismatches very close to the ends of reads, with a strong enrichment of guanine at the right flanking base. That is, an enrichment of CpG-to-TpG mismatches at the ends of reads. This may be explained by the fact that when a methylated cytosine undergoes deamination it becomes thymine (in contrast to unmethylated cytosines, which deaminate to uracil) and these thymines are not repaired by the UDG-treatment [30].

### *3.3.2 The effects of contamination on inferred grades of membership*

We next sought to examine the effects of exogenous modern contamination on inferred grades of memberships in ancient samples. Here, we define contamination as the percentage of reads (with at least one mismatch) that originate from a modern individual.

We performed an *in-silico* experiment to artificially contaminate ancient samples with modern data from the 1000 Genomes Project [1]. We selected one BAM file from an ancient sample (K01 from [42]), and split its reads (with at least one mismatch) into 10 equal subsets. We then contaminated each subset with reads from a distinct modern individual from the 1000 Genomes Project, varying the contamination level from 0% to 10% (Figure 3.3A). This results in 10 samples (S1-S10) representing 10 contaminated ancient samples with known levels of modern contamination.

We applied `aRchaic` with $K = 2$ on the contaminated samples (S1-S10) plus 40 other modern individuals (randomly sampled from the 1000 Genomes Project; Figure 3.3B). Modern individuals showed high grades of membership in one cluster (red). Uncontaminated ancient sample showed essentially no membership in this cluster. For contaminated ancient samples, membership in the red cluster increased linearly with the level of contamination (Figure 3.3-C). We obtained similar results even with only 10000 randomly-sampled reads for each sample (Figure 3.3-D) implying that these results are robust to low sequencing depth.

Although these experiments show that `aRchaic` reflect contamination, the grades of membership may not be direct estimates of the proportion of contamination. Indeed, directly estimating amounts of contamination seems difficult in general because lack of damage does not imply contamination: well-preserved ancient samples may also show reduced levels of DNA damage.

### *3.3.3 `aRchaic` can identify both DNA damage and technical artifacts*

As a final case study, we compiled data from 25 modern and 25 ancient Native Americans from the Northwest Coast of North America [73]. This dataset offers us a opportunity to apply `aRchaic` on modern and ancient DNA samples collected from the same population and sequenced in the same laboratory. In these data, the first two positions from the 5' end of each read had been removed by the original authors in an attempt to mitigate effects of DNA damage. Despite this, `aRchaic`, when applied with $K = 2$ to all 50 samples, clearly distinguishes between modern and ancient individuals (Supplementary Figure S14).

When we applied `aRchaic` to just modern samples we were surprised to find that it also identified two clear clusters (Figure 3.4 panel (a)). These turned out to reflect the fact that the modern samples had been processed using two different library preparation kits, Nextera & True-Seq [73]. Samples prepared with the True-Seq kit showed nearly full membership in one cluster (pink), whereas those prepared with the Nextera kit showed partial membership in a second cluster (beige). These clusters show only small differences in mismatch patterns, but the beige cluster is characterized by a strong excess of mismatches at position 12, apparently an artifact introduced by the Nextera preparation (3.4).

We also applied `aRchaic` with $K = 2$ to just the ancient samples (see Figure 3.4 panel (b)). Unlike the moderns, this yielded a continuous gradient of memberships in the two clusters (blue and gold). One cluster (blue) is dominated by the strong enrichment of $C \rightarrow T$ mismatches relative to the modern background that is typical of ancient samples. The other cluster (gold) is enriched not only for $C \rightarrow T$ mismatches, but also for $C \rightarrow A$ and $T \rightarrow A$ mismatches, possibly representing other types of damage, or other artifacts, in the ancient DNA. Interestingly, the individual with highest membership in this gold cluster was much older than all the others ($\approx$ 6000 years BP; all other samples are $\approx$ 2000 years BP).

## 3.4 Discussion

We developed a method (`aRchaic`) for clustering and visualization of samples based on DNA mismatch patterns. Our method is based on a grade of membership (GoM) model, which generalizes the concept of clustering to allow samples to have membership in multiple clusters. We provide a visual representation of the grades of membership using a "'Structure-plot" [99] and visualization of the mismatch profiles (or clusters) with an *EDLogo* plot [26].

In GoM models, the choice of the number of clusters $K$ (or mismatch profiles) is a contentious issue. In our analyses we selected values of $K$ that highlight interpretable structure in the data. We emphasize that there will typically be no single "true" $K$, and that examining results with different $K$ can often provide additional insights [23]. For example, Figure 2 shows results for $K = 3$, but higher values of $K$ reveal additional structure within each ancient subgroup (Supplementary Figure S12).

A key challenge in analyzing ancient DNA is that data are often contaminated with exogenous modern DNA. Several approaches have been suggested to estimate the amount of contamination. One approach is to compute the rate of polymorphism across the X chromosome in males [65, 94], where the presence of polymorphism would suggest contamination because males have only one X chromosome. Another approach is to quantify the contribution of a panel of modern mitochondrial haplotypes to the ancient DNA [41, 95]. Both of these approaches require reasonably high sequencing depth. `aRchaic` is not an explicit model of contamination, but in some settings (e.g. Figure 3.3) the inferred grades of membership can reflect relative degrees of contamination even with low sequencing depth, and may be a useful complement to these other methods.

Here we have chosen to model features at the level of mismatches which have been shown to be informative of DNA damage in previous studies [16, 45, 60]. Alternatively, one could formulate a model at the level of reads. For example, the method *PMDtools* computes a score for every read representing the probability that the read is damaged [117]. This method models mismatches along the read; additionally, one can incorporate indels and fragment length along with mismatches. One reason we chose not to model these extra features was to reduce the feature and computational complexity of our method. Furthermore, these extra features may not actually be driven by DNA damage. For example, we explored fragment length profiles in several aDNA data-sets and found their distributions to be primarily driven by lab-specific effects rather than DNA damage. Another limitation of fragment length is that it can be used only in studies using paired-end sequencing.

Methods described here are available in a R/python open-source software package at `www.github.com/kkdey/aRchaic`.

## 3.5 Author Contributions

Dey, KK and Al-Asadi, H designed the method. Dey, KK and Al-Asadi, H implemented the method. Dey, KK and Al-Asadi, H ran the experiments. Dey, KK and Al-Asadi, H produced the figures. Dey, KK, Al-Asadi H, and Stephens, M wrote the paper.

## 3.6 Figures

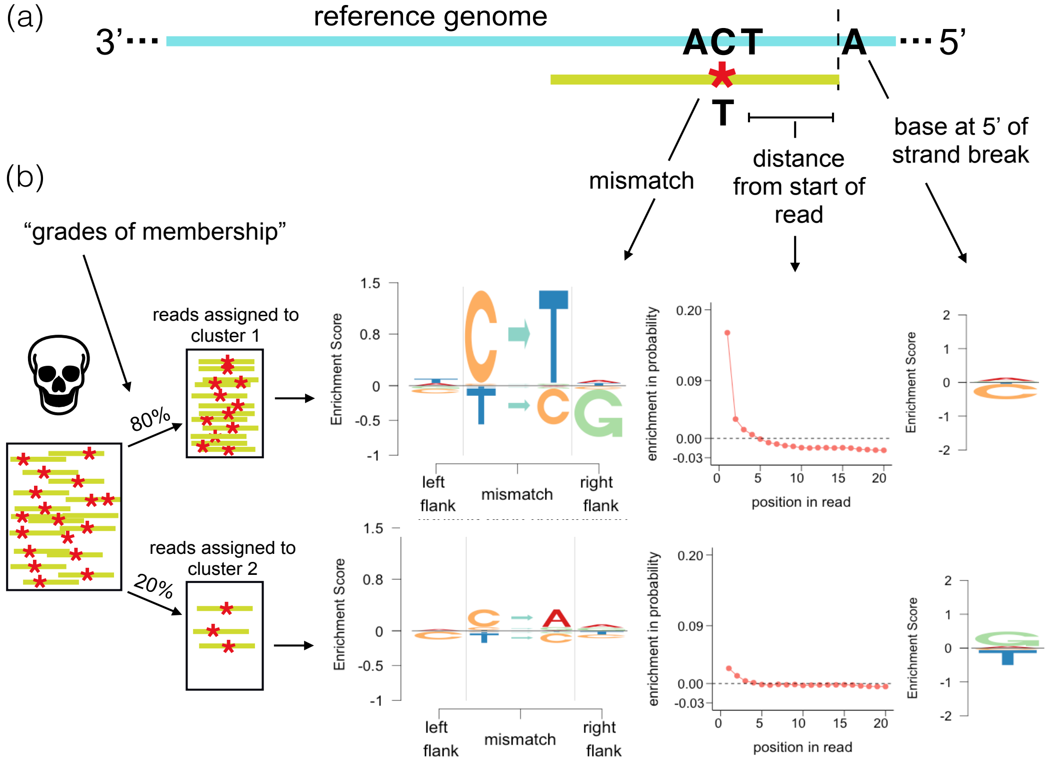


Figure 3.1: Illustration of the `aRchaic` grades of membership and mismatch profiles. (a) The features of a mismatch modeled by `aRchaic` (b) A depiction of an ancient DNA sample that has 80% of it's reads assigned to cluster 1 cluster and 20% of it's reads assigned to cluster 2. Each cluster is defined by a *mismatch profile* showing the enrichment of the mismatch type, bases flanking the mismatch, the distance of the mismatch from the nearest end of the read, and the base immediately 5' to the strand-break. To produce a mismatch profile for a cluster, mismatch features are aggregated across reads assigned to the cluster, and their frequencies are represented by an *EDLogo* plot [26]. In the *EDLogo* plot, the frequencies are scaled against a background frequency computed from The 1000 Genomes Project [1].

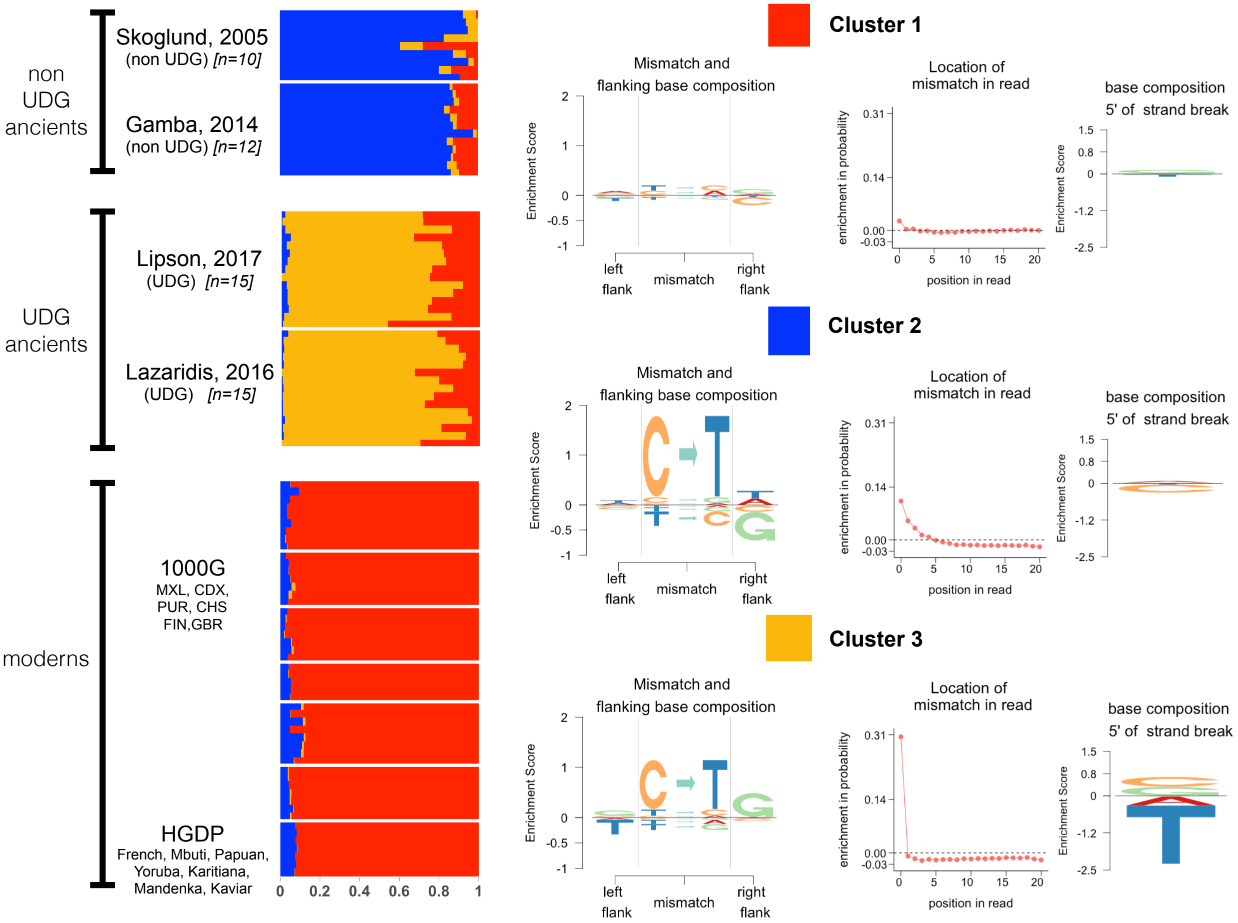


Figure 3.2: `aRchaic` clearly distinguishes between modern, ancient (UDG), and ancient (non-UDG) samples `aRchaic` is applied with $K = 3$ to a collection of ancient individuals from four studies [42, 68, 79, 116] along with modern individuals randomly sampled from the 1000 Genomes Project and the Human Genome Diversity Panel [1, 17]. Modern samples have high membership in the red cluster. The *EDLogo* representation of this cluster does not show strong enrichment against a modern background. The ancient (non-UDG) samples are representative of the blue cluster. The *EDLogo* plot for the blue cluster shows a strong enrichment in $C \rightarrow T$ mismatches at the end of reads, a depletion of Guanine in the right flanking base, and a depletion of Cytosine at the 5' strand-break. The ancient (UDG) samples have partial membership both in the red cluster and in the gold cluster. The *EDLogo* plot for the gold cluster is enriched in $C \rightarrow T$ mismatches at the terminal ends of the reads, and also shows an enrichment of Guanine at the right flanking base.

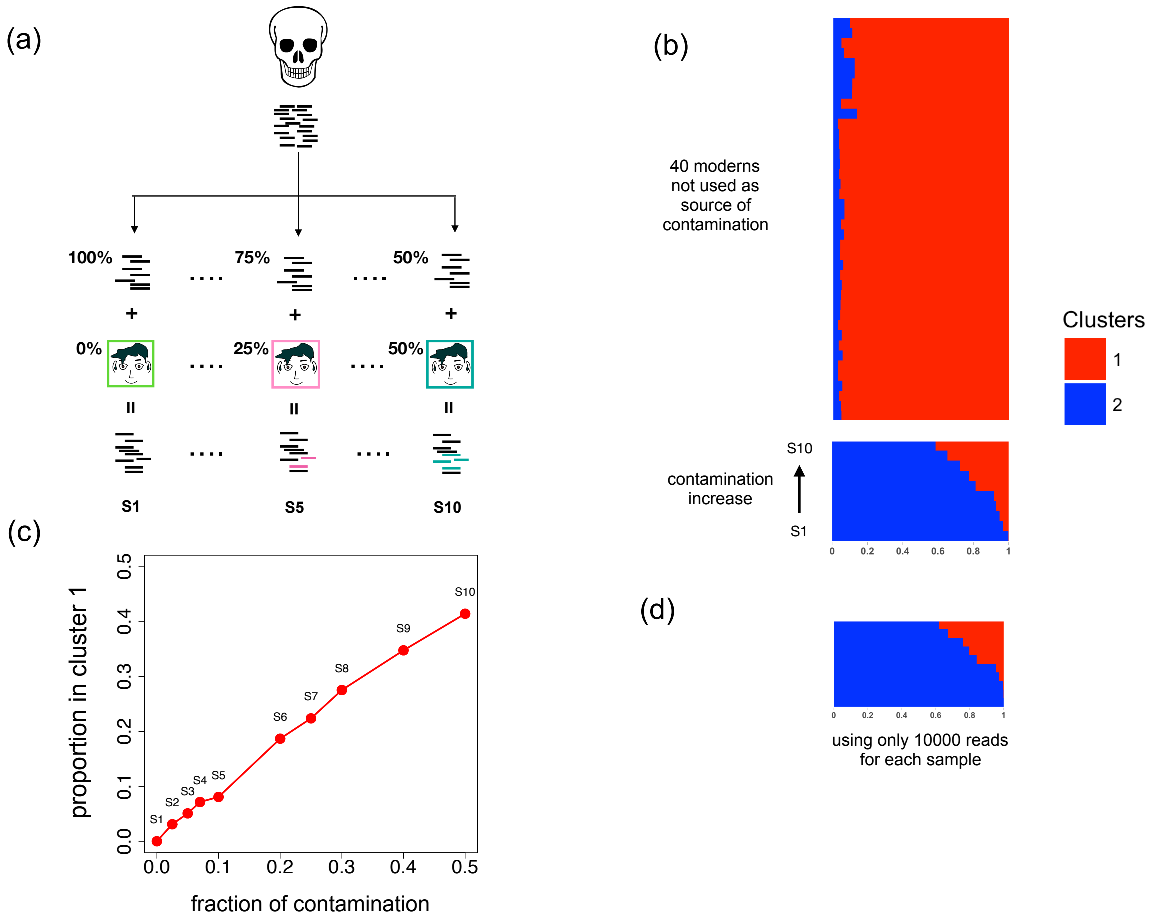


Figure 3.3: Estimated grades of membership reflect levels of contamination (a) Reads from one ancient individual (KO1 from [42]) were split into 10 equally sized groups. To each of these groups, reads (with at least one mismatch) were added from a distinct individual in the 1000 Genomes Project [1] to each group (S1-S10) at varying levels of contamination. (b) We applied `aRchaic` with $K = 2$ on a combined dataset comprised of these 10 contaminated groups of reads (S1-S10) along with 40 other modern individuals from 1000 Genomes (c) The grades of membership in cluster (red) were plotted as a function of the percentage of contamination. (d) Each group (S1-S10) was further sub-sampled to 10,000 reads, and `aRchaic` was applied with $K = 2$ to the new subsampled groups and the same 40 modern individuals as in panel b.

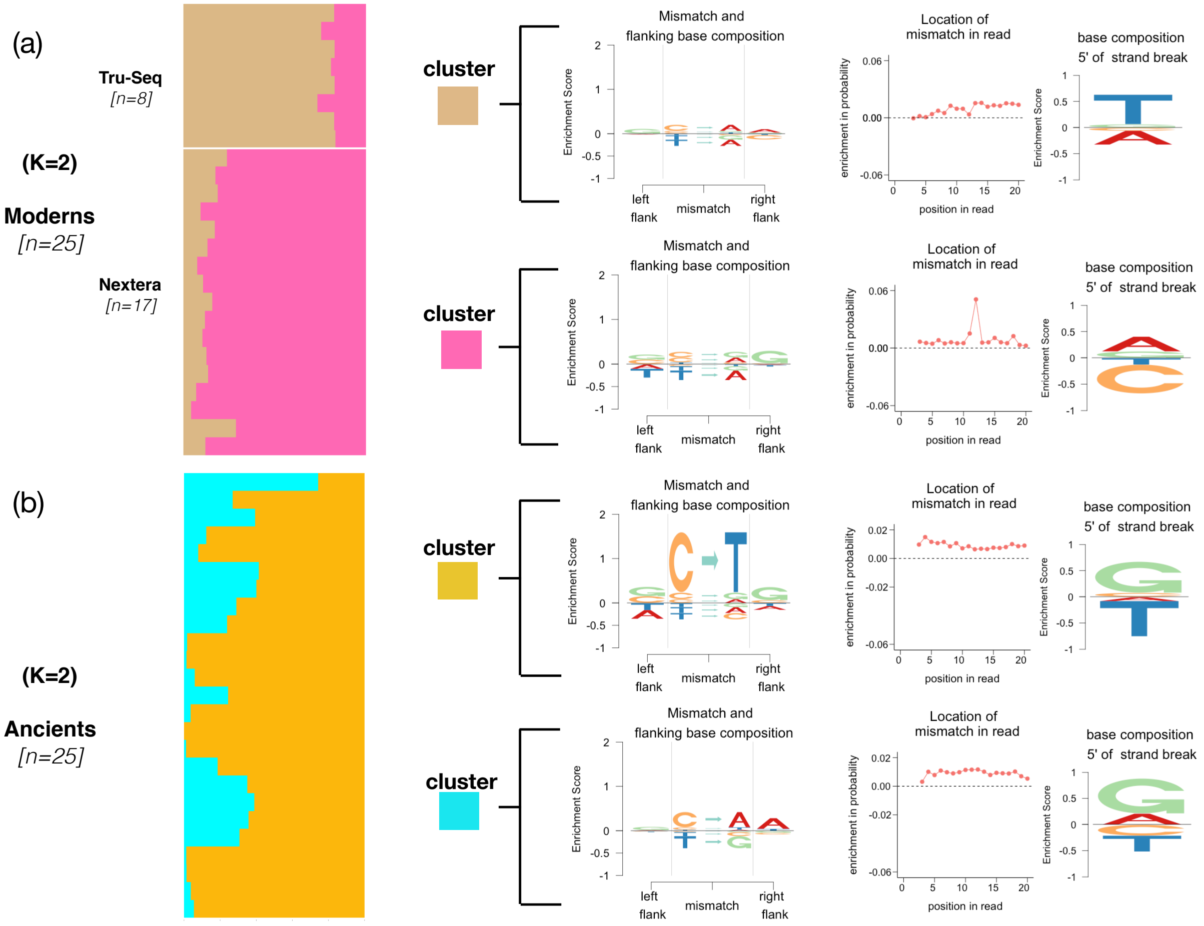


Figure 3.4: DNA damage and library preparation techniques drive grades of membership (a) We applied aRchaic with $K = 2$ to 25 modern samples from [73]. The samples prepared with the True-Seq kit show nearly full membership in the pink cluster. Samples prepared with the Nextera kit show partial membership in the pink cluster and the tan cluster. The tan cluster shows a blip at the 12th position from the end of the read (b) We applied aRchaic with $K = 2$ to 25 ancient samples from [73]. The two clusters show an enrichment of $C \to T$ mismatches at the ends of reads and an enrichment of purines at the 5' strand-break.

# Chapter 4

# *Logolas* : Enrichment Depletion and String Logo Plots

*(with Dongue Xie and M. Stephens)*

## 4.1 Introduction

Since their introduction in the early 90's by Schneider and Stephens [109], sequence logo plots have become widely used for visualizing short conserved patterns known as *sequence motifs*, in multiple alignments of DNA, RNA and protein sequences. At each position in the alignment, the standard logo plot represents the relative frequency of each character (base, amino acid etc) by stacking characters on top of each other, with the height of each character proportional to its relative frequency. The characters are ordered by their relative frequency, and the total height of the stack is determined by the information content of the position. The visualization is so appealing that methods to produce logo plots are now implemented in many software packages (e.g. *seqLogo* [8], *RWebLogo* [130], *ggseqlogo* [131]) and web servers (e.g. *WebLogo* [20], *Seq2Logo* [125], *iceLogo* [18]).

Because the standard logo plot scales the height of each character proportional to its relative frequency, it tends to visually highlight characters that are *enriched*; that is, at higher than expected frequency. In many applications such enrichments may be the main features of interest, and the standard logo plot serves these applications well. However, sometimes it may be equally interesting to identify *depletions*: characters that occur *less often* than expected. The standard logo plot represents strong depletion by the *absence* of a character, which produces less visual emphasis than an enrichment.

To better highlight depletions in amino acid motifs [125] suggest several alternatives to the standard logo plot. The key idea is to explicitly represent depletions using characters that occupy the negative part of the $y$ axis. However, we have found that the resulting plots sometimes suffer from visual clutter – too many symbols, which distract from the main patterns of enrichment and depletion.

Here we suggest a simple solution to this problem, producing a new sequence logo plot – the *Enrichment Depletion Logo* or *EDLogo* plot – that highlights both enrichment and depletion, while minimizing visual clutter. In addition, we extend the applicability of logo plots to new settings by i) allowing each "character" in the plot to be an arbitrary alphanumeric string (potentially including user-defined symbols); and ii) allowing a different "alphabet" of permitted strings at each position. All these new features are implemented in our R package, *Logolas*, which can produce generalized string-based logo and *EDLogo* plots. We illustrate the utility of the *EDLogo* plot and the flexibility of the string-based representation through several applications.

## 4.2 Implementation

### *4.2.1 Intuition*

In essence, the goal of a logo plot is to represent, at each position along the $x$ axis, how a probability vector $\mathbf{p}$ compares with another probability vector $\mathbf{q}$. For example, suppose that at a specific position in a set of aligned DNA sequences, we observe relative frequencies $\mathbf{p} = (p_A, p_C, p_G, p_T) = (0.33, 0.33, 0.33, 0.01)$ of the four bases $\{A, C, G, T\}$. The goal of the logo plot might be to represent how $\mathbf{p}$ compares with the background frequencies of the four bases, which for simplicity we will assume in this example to be equal: $\mathbf{q} = (q_A, q_C, q_G, q_T) = (0.25, 0.25, 0.25, 0.25)$. Verbally we could describe the change from $\mathbf{q}$ to $\mathbf{p}$ in several ways: we could say "$T$ is depleted", or "$A, C$ and $G$ are enriched", or "$T$ is depleted, and $A, C$ and $G$ are enriched". While all of these are valid statements, the first is the most succinct, and our *EDLogo* plot provides a visual version of that statement. The second statement is more in line with a standard logo representation, and the last is in essence the approach in [125]. See Figure 4.1.

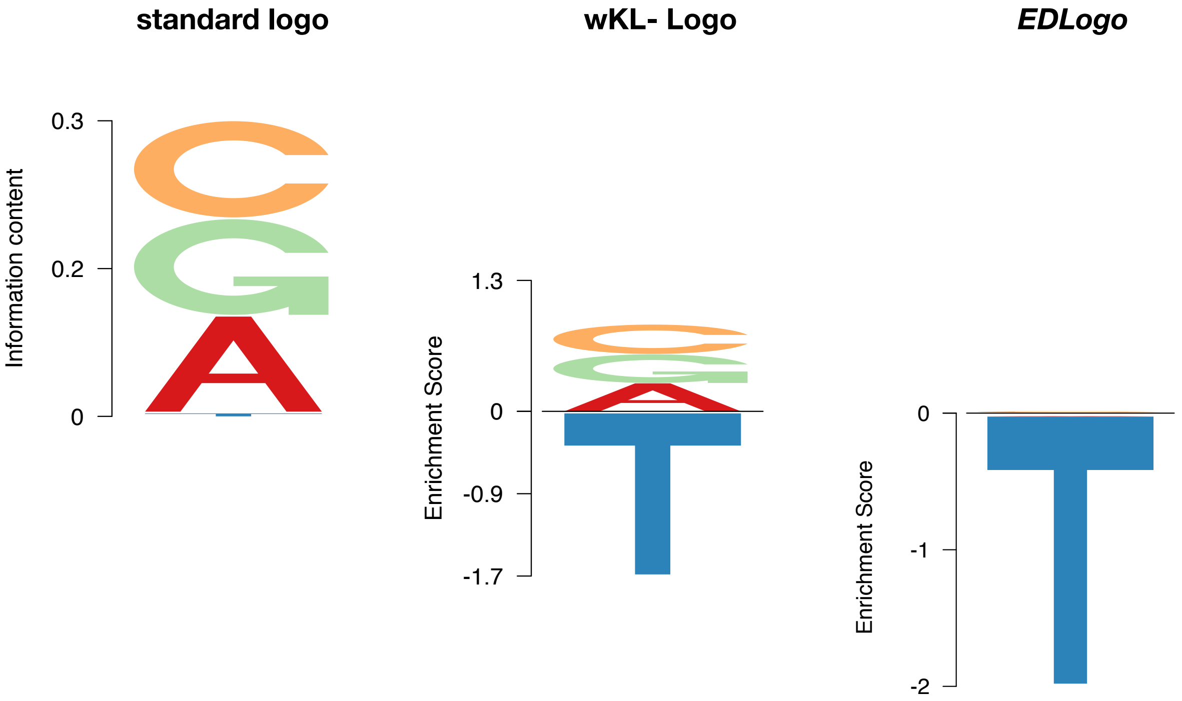


Figure 4.1: **Illustration of the differences between standard logo, *EDLogo* and wKL-Logo representations.** The figure shows how the different logos represent observed frequencies $\mathbf{p} = (p_A, p_C, p_G, p_T) = (0.33, 0.33, 0.33, 0.01)$ (compared with a uniform background, $\mathbf{q} = (0.25, 0.25, 0.25, 0.25)$). The standard logo effectively represents $\mathbf{p}$ by highlighting that “A, C and G are enriched”; *EDLogo* represents it by highlighting “T is depleted”; wKL-Logo represents it as “A, C and G are enriched and T is depleted”. All are correct statements, but the *EDLogo* representation is the most parsimonious.

### *4.2.2 The EDLogo plot*

At a particular position, $j$, of a sequence (or other indexing set), let $\mathbf{p} = (p_1, p_2, \ldots, p_n)$ denote the probabilities of the $n$ elements $C_1, \ldots, C_n$ (which can be characters or strings) permitted at that position, and $\mathbf{q} = (q_1, q_2, \ldots, q_n)$ denote corresponding background probabilities. Define $\mathbf{r} = (r_1, r_2, \ldots, r_n)$ by:

$$r_i = \log_2 \frac{p_i}{q_i} - \text{median}\left(\left\{\log_2 \frac{p_i}{q_i} : i = 1, 2, \ldots, n\right\}\right). \tag{4.1}$$

Then at position $j$ along the $x$ axis, the *EDLogo* plot plots the element $C_i$, scaled to have height $|r_i|$, and above the $x$ axis if $r_i$ is positive, or below the $x$ axis if $r_i$ is negative. Elements are stacked (from bottom to top) in order of increasing $r_i$, so that the largest characters are furthest from the axis. (In practice, to avoid potential numerical issues if $p_i$ or $q_i$ are very small, we add a small value $\epsilon$ to each element $p_i$ and $q_i$ before computing $r_i$; default $\epsilon = 0.01$.)

The basic strategy has close connections to ideas in [125], but with the crucial difference that we subtract the median in Equation 4.1. As our examples will demonstrate, subtracting the median in this way – which can be motivated by a parsimony argument (see below) – can dramatically change the plot, and substantially reduce visual clutter.

Note that the *EDLogo* plot for $\mathbf{p}$ vs $\mathbf{q}$ is essentially a mirror (about the $x$ axis) of the *EDLogo* plot for $\mathbf{q}$ vs $\mathbf{p}$ (e.g. Supplementary Figure S16). We call this the "mirror property", and it can be interpreted as meaning that the plots treat enrichment and depletion symmetrically. This property is also satisfied by plots in [125], but not by the standard logo plot.

## A model-based view

Suppose we model the relationship of $\mathbf{p}$ to $\mathbf{q}$ by

$$p_i \propto \lambda_i q_i \tag{4.2}$$

for some unknown (positive) "parameters" $\lambda_i$. For example, this model would arise if $\mathbf{q}$ represents the underlying frequencies of elements in a population, and $\mathbf{p}$ represents the frequencies of the same elements in a (large) sample from that population, conditional on an event $E$ (e.g. a transcription factor binding). Indeed, by Bayes theorem, under this assumption we would have

$$p_i \propto \Pr(E|\text{element } i) q_i. \tag{4.3}$$

Since the $p_i$ must sum to 1, $\sum_i p_i = 1$, the model (4.2) implies

$$p_i = \lambda_i q_i / \sum_j \lambda_j q_j. \tag{4.4}$$

Now consider estimating the parameters $\lambda$. Even if $\mathbf{p}$ and $\mathbf{q}$ are observed without error, there is a non-identifiability in estimating $\lambda$: we can set $\lambda_i = c p_i / q_i$ for any positive $c$. Equivalently, if we consider estimating the logarithms $l_i := \log \lambda_i$, we can set

$$l_i = \log_2(p_i/q_i) + k \tag{4.5}$$

for any constant $k$. Note that $r_i$ in (4.1) has exactly this form, and so the vector $\mathbf{r}$ can be interpreted as an estimate of the vector $\mathbf{l}$. Furthermore, it is easy to show that, among all estimates of the form (4.5), $\mathbf{r}$ has the smallest sum of absolute values. That is, $\mathbf{r}$ solves the optimization

$$\mathbf{r} = \arg\min_{\mathbf{l}} \sum_i |l_i| \tag{4.6}$$

subject to the constraint (4.5).

Since the sum of absolute values of $\mathbf{r}$ is the total height of the stacked characters in the *EDLogo* plot, one can think of our choice of $\mathbf{r}$ as the *estimate of* $\mathbf{l}$ *that produces the smallest stack of characters* – that is, the most “parsimonious” estimate.

### 4.2.3 Interpretation

Roughly speaking, positive values of $r_i$ can be interpreted as indicating characters that are "enriched" and negative values of $r_i$ as indicating characters that are "depleted". Formally we must add that here enrichment and depletion are to be interpreted as *relative to the median enrichment/depletion across characters.* This relative enrichment does not necessarily imply enrichment or depletion in some "absolute" sense: for example, $r_i$ could be positive even if $p_i$ is smaller than $q_i$. For compositional data it seems natural that enrichment/depletion be interpreted relative to some "baseline", and our choice of the median as the baseline is motivated above as providing the most parsimonious plot.

It may also help interpretation to note that for any two characters $i$ and $i'$, the difference $r_i - r_{i'}$ is equal to the log-odds ratio:

$$r_i - r_{i'} = \log_2\left(\frac{p_i/p_{i'}}{q_i/q_{i'}}\right). \tag{4.7}$$

### 4.2.4 A variation: the scaled EDLogo plot

In the standard logo plot the total height of the stack at each position is scaled to reflect the "information content" at that position, or, more generally, the Kullback–Leibler divergence (KLD) from the background frequencies $\mathbf{q}$ to the observed frequencies $\mathbf{p}$ [120]. This scaling highlights locations where $\mathbf{p}$ differs most strongly from $\mathbf{q}$. Similarly, the stack heights in the *EDLogo* plot also reflect the extent to which $\mathbf{p}$ differs from $\mathbf{q}$; for example, if $\mathbf{p} = \mathbf{q}$ then the stack height is 0. However, the *EDLogo* stack heights are not equal to the KLD.

Empirically, compared with the standard KLD stack heights, the stack heights in the *EDLogo* plot tend to down-weight locations with a single strongly-enriched element. In settings where this is undesirable, we could avoid it by scaling the *EDLogo* plot to match the standard plot. That is, we could scale the elements at each position by a (position-specific) constant factor so that the stack height is, like the standard plot, equal to the KLD. However, this would lose the mirror property of the *EDLogo* plot because the KLD is not symmetric in **p** and **q**. Thus we instead suggesting scaling by the symmetric KL divergence (symmKLD) between **p** and **q**, which highlights strong single-element enrichments while retaining the mirror property. We call the resulting plot the *scaled EDLogo* plot.

## 4.3 Results

### *4.3.1 Comparison with existing logo plots*

Figure 4.2 illustrates the *EDLogo* plot, and compares it with the standard logo and the weighted Kullback–Leibler logo (wKL-Logo) plot [125], in four diverse applications.

The first two applications (panels (a) and (b)) are settings where the standard logo plot is widely used: visualizing transcription factor binding sites (TFBS) [59, 63, 103, 122, 135, 139], and protein binding motifs [61, 111]. These examples showcase the effectiveness of the standard logo plot in highlighting enrichments: in our opinion it does this better than the other two plots, and in this sense the other plots should be viewed as complementing the standard plot rather than replacing it. These examples also illustrate the differences between the wKL-Logo and *EDLogo* plots, both of which aim to highlight depletion as well as enrichment: the *EDLogo* plot introduces less distracting visual clutter than the wKL-Logo plot, producing a cleaner and more parsimonious visualization that better highlights the primary enrichments and depletions. In particular, for the TFBS example (panel (a), which shows the primary discovered motif *disc1* of Early B cell factor EBF1 from ENCODE [63]), the *EDLogo* plot is most effective at highlighting depletion of bases G and C at the two positions in the middle of the sequence. This depletion is hard to see in the standard logo because of its emphasis on enrichment, and less clear in the wKL-Logo due to visual clutter. This depletion pattern is likely meaningful, rather than a coincidence, since it was also observed in two other previously known motifs of the same transcription factor [59, 103]

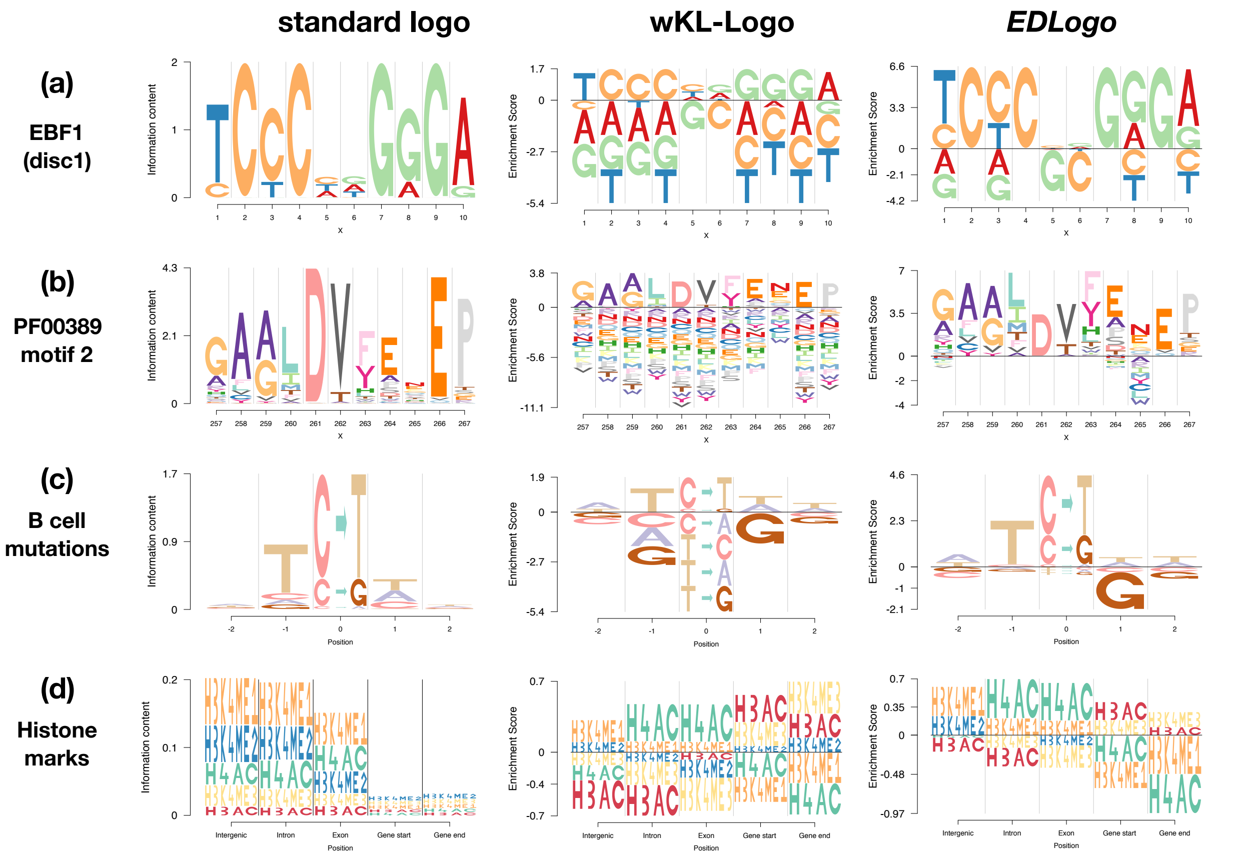


Figure 4.2: **Comparison of standard logo plot, weighted KL (w-KL) logo plot and *EDLogo* plot on four examples.** Panel (a): the transcription factor binding site of the EBF1-disc1 transcription factor. Panel (b): the binding motif (Motif2 Start=257 Length=11) of the protein *D-isomer specific 2-hydroxyacid dehydrogenase, catalytic domain (IPR006139)* from [61, 111]. Panel (c): mutational signature profile of mutations in lymphoma B cells, with data from [4]. The depletion of G to the right of the mutation - possibly occurring due to the rarity of CpG sites owing to de-amination of methylated cytosines - is clearest in the *EDLogo* representation. Panel (d): relative abundance of histone modification sites across various genomic regions in the lymphoblastoid cell line GM06990 (Table S2 in Koch et al 2007 [64]). These examples illustrate the ability of the *EDLogo* plot to highlight both enrichment and depletion, while avoiding unnecessarily visual clutter. The last two examples also illustrate how our software allows arbitrary strings as elements in a logo plot.

(see Supplementary Figure S17).

The next two applications (panels (c) and (d) of Figure 4.2) are non-standard settings that illustrate the use of general strings as "characters" in a logo plot, as well as providing further examples where the *EDLogo* plot is particularly effective at highlighting depletion as well as enrichment.

Panel (c) shows logo plots representing a cancer mutational signature from lymphoma B cell somatic mutations [4]. Here we follow [115] in representing a mutational signature by the frequency of each type of mutation, together with base frequencies at the ±2 flanking bases. We also follow the common convention of orienting the strand so that the mutation is from either a $C$ or a $T$, yielding six possible mutation types: $C \to T$, $C \to A$, $C \to G$, $T \to A$, $T \to C$, $T \to G$. This Figure panel illustrates two important points. First, it illustrates the flexibility of our software package *Logolas*, which allows arbitrary strings in a logo. For all three logo plots (standard, wKL and ED) we use this to represent the six mutation types by six strings of the form $X \to Y$, and we find the resulting plots easier to read than the *pmsignature* plots in [115] (see Supplementary Figure S18 for comparison). Additionally, it also shows that one can use different sets of permitted strings at different positions - strings are only used to represent the mutation in the center, while characters are used to represent the flanking bases. Second, it illustrates a case where, in our opinion, the *EDLogo* plot is a better visual summary than the other plots. Specifically the *EDLogo* plot best highlights the three primary aspects of this signature: enrichment of $C \to T$ and $C \to G$ mutation types; enrichment of $T$ at position -1; and depletion of $G$ at position +1. Here the depletion of $G$ at +1 may be a bi-product of the enrichment of $C \to \cdot$ mutation types combined with the overall depletion of CpG sites in the genome due to deamination [105]. For readers interested in other cancer mutation signatures, we provide *EDLogo* plots for 24 cancer mutation signatures from [4] in Supplementary Figure S19.

Panel (d) shows logo plots summarizing the *relative* abundance of 5 different histone marks in different genomic contexts (data from lymphoblastoid cell line GM06990, Table S2 (*upper*) of [64]; background probabilities from Table S2 (*lower*) of [64]). Note that relative abundances yield compositional data that can be visualized in a logo plot. Again this example illustrates the potential to use strings in logo plots. It also represents an example where the *EDLogo* and wKL-Logo plots seem more informative than the standard logo plot. Specifically, the standard logo plot is dominated by the high deviation from background frequencies at the intergenic, exon and intron regions, and the differences in enrichments and depletions among regions are difficult to discern. In comparison, the *EDLogo* and wKL-Logo plots highlight a number of differences among regions (some of which are also noted in [64]). For example, both plots highlight the relative enrichment of H3AC and H3K4me3 near the start and end of genes, and corresponding relative depletion of H4AC and H3K4me1. Both plots also highlight relative enrichment of H3K4me1 compared with other marks in the intergenic, exonic and intronic regions; the relative enrichment of H4AC in intronic and exonic regions, and relative depletion of H3AC in intergenic and intronic regions.

### *4.3.2 The scaled EDLogo plot*

In the first two applications above (panels (a) and (b) of Figure 4.2) we noted the effectiveness of the standard logo plot in highlighting strong enrichments. This stems from its use of the KLD to scale stack heights at each position. Motivated by this, we implemented a *scaled EDLogo* plot, which combines properties of the *EDLogo* plot (highlighting both enrichments and depletions) and the standard plot (scaling stack heights based on KLD). The *scaled EDLogo* plot for all four of the examples in Figure 4.2 are shown in Supplementary Figure S20. The results – particularly panels (a) and (b) – illustrate how the *scaled EDLogo* plot tends to emphasize strong enrichments more than the unscaled version, so the scaled version may be preferred in settings where such enrichments are the primary focus.

### *4.3.3 Further Variations*

Further variations on the *EDLogo* plot can be created by replacing $\log_2(p_i/q_i)$, in (4.1) with other functions of $(p_i, q_i)$, such as the log-odds, $\log_2(p_i/(1-p_i)) - \log_2(q_i/(1-q_i))$. We have not found any particular advantage of such variations over the *EDLogo* plot presented here, but several such variations are implemented in the software and also illustrated in Supplementary Figure S21. In addition, the *EDLogo* strategy of using a median adjustment in (4.1) to reduce visual clutter can be directly applied to derived quantities such as the position specific scoring matrix (PSSM) [61, 111], commonly used to represent protein binding motifs (Supplementary figure S22).

## 4.4 Discussion

We present a new sequence logo plot, the *EDLogo* plot, designed to highlight both enrichment and depletion of elements at each position in a sequence (or other index set). We have implemented this plot, as well as standard logo plots, in a flexible R package *Logolas*, which offers many other features: the ability to use strings instead of characters; various customizable styles and color palettes; several methods for scaling stack heights; and ease of integrating logo plots with external graphics like ggplot2 [134].

The Logolas R package is currently under active development on Github (`https://github.com/kkdey/Logolas`). Code for reproducing figures in this paper is available at `https://github.com/kkdey/Logolas-paper`. Vignettes and a gallery demonstrating features of Logolas are available at (`https://github.com/kkdey/Logolas-pages`)

## 4.5 Author contributions

Dey, KK and Stephens, M designed the method. Dey, KK implemented the method. Dey, KK and Xie, D ran the experiments. Dey, KK and Xie, D produced the figures. Dey, KK and Stephens, M wrote the paper.

## 4.6 Supplementary Methods

Here we detail several alternative options we have implemented for computing the values of $r_i$ when creating an *EDLogo* plot to compare observed relative frequencies $\mathbf{p}$ with background frequencies $\mathbf{q}$:

- *log ratio* approach

$$r_i = \log_2 \frac{p_i + \epsilon}{q_i + \epsilon} - \text{median}\left(\left\{\log_2 \frac{p_i + \epsilon}{q_i + \epsilon} : i = 1, 2, \ldots, n\right\}\right) \tag{4.8}$$

- *log-odds ratio* approach

$$r_i = \log_2 \frac{p_i/(1-p_i) + \epsilon}{q_i/(1-q_i) + \epsilon} - \text{median}\left(\left\{\log_2 \frac{p_i/(1-p_i) + \epsilon}{q_i/(1-q_i) + \epsilon} : i = 1, 2, \ldots, n\right\}\right) \tag{4.9}$$

- *ratio* approach

$$r_i = \frac{p_i + \epsilon}{q_i + \epsilon} - \text{median}\left(\left\{\frac{p_i + \epsilon}{q_i + \epsilon} : i = 1, 2, \ldots, n\right\}\right) \tag{4.10}$$

- *probKL* approach [125]

$$r_i = p_i \log_2 \frac{p_i + \epsilon}{q_i + \epsilon} - \text{median}\left(\left\{p_i \log_2 \frac{p_i + \epsilon}{q_i + \epsilon} : i = 1, 2, \ldots, n\right\}\right) \tag{4.11}$$

The *log ratio* approach is our default choice and is the one discussed in detail in the main text. Just like the *log ratio* approach, each of the other options also has its corresponding *scaled* version as demonstrated in Supplementary Figure S21.

# Chapter 5

# *CorShrink*: Adaptively Parsimonious Representation of Correlation Matrices

*(with M. Stephens)*

## 5.1 Introduction

Estimating the correlation matrix of a set of variables is one of the fundamental problems in statistics. The standard estimator, the sample correlation matrix, is not efficient under certain settings. One such setting is when the dimensionality of the problem ($p$) considerably exceeds the number of samples ($n$). This problem has driven statisticians to suggest various alternatives - for example - convex combination of sample correlation with one or more target correlation matrices [66, 69, 70, 107, 126], optimal thresholding of correlations [10, 102] or LASSO type shrinkage on correlation or inverse correlation matrices [11, 40]. These methods, though effective, have their own sets of limitations. For example, the thresholding methods [10, 102] work under specific assumptions of bandedness on the population correlation matrix, while the LASSO-based models [11, 40] often require extensive cross validation to tune the shrinkage parameter.

Another setting where the sample correlation matrix is an inefficient estimator is when there are large scale missing observations in the underlying data matrix. In this case, the correlations between different pairs of variables are computed over different numbers of *matched samples* - samples that have observations recorded for both variables of the pair. Consequently, correlations computed over a small number of matched samples are typically less trustworthy. One solution to this problem is to impute the missing values in the data matrix [80] and then estimate the correlations based on the imputed data. However, as we show below, this approach is subject to imputation error and can create substantial biases in case of large scale missing data.

Here, we propose a fast simplistic approach called `CorShrink` that adaptively shrinks the correlation matrix without requiring the user to perform cross validation and works under minimal assumptions on population correlation structure. We perform simulation studies to show that under *small n, large p* settings and sparse structure assumption on the population correlation matrix, `CorShrink` outperforms other popularly used methods of correlation shrinkage. Also, `CorShrink` is flexible in handling data matrices with missing observations and accounts for the differences in the number of matched samples between variables in a model based way. As an example application, `CorShrink` is applied to the donor by tissue expression data matrix for a gene in the Genotype Tissue Expression (GTEx) project [75], where the data contains large scale missing observations owing to each donor contributing only a few tissues. The estimated correlation matrix using `CorShrink` is found to be less visually cluttered and more interpretable than the corresponding pairwise sample correlation matrix. Furthermore, the modeling approach of `CorShrink` extends beyond sample correlations to other correlation-like quantities such as cosine similarities of vectors and can be used to generate more accurate similarity measures and word-word similarity rankings from *word2vec* models [82, 83].

## 5.2 Methods

Let $(X_{np})_{N \times P}$ denote a data matrix with $N$ samples and $P$ variables, where some values may be missing (recorded as NA). For each pair of variables $i, j \in \{1, 2, \cdots, P\}$ let $R_{ij}$ denote their (unknown) true correlation, and $\hat{R}_{ij}$ denote the sample correlation computed using only the samples $n$ that have observed values for both the variables $i$ and $j$ (e.g. using the option `use="pairwise.complete.obs", method = "pearson"` in the R function `cor`).

Further, let $Z_{ij}$ and $\hat{Z}_{ij}$ denote the corresponding Fisher Z-transforms [38]:

$$Z_{ij} = Z(R_{ij}) = \frac{1}{2}\log\left(\frac{1+R_{ij}}{1-R_{ij}}\right) \tag{5.1}$$

$$\hat{Z}_{ij} = Z(\hat{R}_{ij}). \tag{5.2}$$

Under a bivariate normality assumption on each pair of variables, [39] showed that the observations $\hat{Z}_{ij}$ are approximately normal:

$$\hat{Z}_{ij}|Z_{ij} \sim N(Z_{ij}, s_{ij}) \tag{5.3}$$

with standard error

$$s_{ij} = \sqrt{\frac{1}{(n_{ij}-1)} + \frac{2}{(n_{ij}-1)^2}}, \tag{5.4}$$

where $n_{ij} > 3$ is the number of *matched samples*, for which both variables $i$ and $j$ are observed:

$$n_{ij} := \#\left\{n : \; X_{ni} \neq \text{NA}, \;\; X_{nj} \neq \text{NA}\right\}. \tag{5.5}$$

We assume a composite likelihood for these Z-scores $Z_{ij}$.

$$L \propto \prod_{i \neq j, i=1(|)P, j=1(|)P} N\left(Z_{ij}|\eta_{ij}, s_{ij}\right) \tag{5.6}$$

where $\eta_{ij}$ is the population Z-score parameter between variables $i$ and $j$. Under the `CorShrink` model, we assume an unimodal mixture model prior for $\eta$ in Equation 5.7.

$$\Pi(\eta_{ij}) := \sum_{k=1}^{K} \pi_k F_k\left(\eta_{ij}\right) \tag{5.7}$$

where $F_k$'s represent component distributions in the mixture, that are empirically determined from the data and the $\pi_k$ are the mixing parameters that are estimated by fitting the model.

We consider three major class of distributions from which the $F_k$ distributions are drawn.

- *Normal* : $F_k = N\left(0, \sigma_k^2\right)$
- *Uniform* : $F_k = U\left(-a_k, a_k\right)$
- *Half Uniform* : $F_k = U\left(0, a_k\right)$ and/or $U\left(-a_k, 0\right)$
- *Nonparametric* : $F_k = U\left(b_k, b_{k+1}\right)$ for $k = 1, 2, \cdots, K$ with $b_1 < b_2 \cdots < b_k < b_{k+1} < \cdots b_{K+1}$ and $b_{k+1} - b_k = c$, with $K \times c$ greater than the range of the Z scores.

This modeling framework heavily draws from the adaptive shrinkage (*ash*) method developed by one of the authors, M. Stephens [119], for shrinking effect sizes in calculating false discovery rates.

The first three choices of $F$ (normal, uniform and half-uniform) can be used to approximate any unimodal distribution centered around 0, using a sufficiently large grid of $\sigma_k$ or $a_k$. For practical purposes, The values $\sigma_k$ or $a_k$ are empirically determined based on the range of the Fisher's Z-score values. One can also specify a background mode apart from 0 to center the component distributions. The nonparametric choice of $F$ on the other hand can approximate any distribution for sufficiently small $c$, thereby further increasing the flexibility of the prior.

The above model (likelihood: Equation 5.6, prior : Equation 5.7) is fitted to obtain the posterior mean of $\eta_{ij}$, $\eta^{\star}_{ij}$, given $R_{ij}$.

$$\eta^{\star}_{ij} := E\left[\eta_{ij} | R_{ij}\right] \tag{5.8}$$

$\eta^{\star}_{ij}$ are adaptively shrunk estimates of the Fisher Z-scores $Z_{ij}$ that account for $n_{ij}$, the number of matched samples between variables $i$ and $j$. The smaller the $n_{ij}$, the higher would be $s_{ij}$ in Equation 5.3 and higher would be the level of shrinkage on $Z_{ij}$.

Next, we reverse transform the Z-scores to get back shrunk estimates of correlation ($r^{\star}_{ij}$).

$$r^{\star}_{ij} := \frac{exp(2\eta^{\star}_{ij}) - 1}{exp(2\eta^{\star}_{ij}) + 1} \tag{5.9}$$

The matrix $R^{\star} = ((r^{\star}_{ij}))_{P \times P}$ may not be positive definite. So, we select the nearest positive definite matrix $R^{\star\star}$ to $R^{\star}$, using the method from [53].

If the variables are not pairwise normally distributed, then the representation of $s_{ij}$ as per Equation 5.3 does not hold. One approach in this context is to use transformations of the data that are more robust to the non-normality of the data, for example - Box-cox, ranks, rank-based inverse normal (RIN) transformations [12, 13]. Another approach would be to estimate $s_{ij}$ using Bootstrapping [28, 31] on the samples. The flexibility to use re-sampling methods as above, extends the scope of the `CorShrink` method beyond correlations to any correlation-like quantities -partial correlations, rank correlations, cosine similarities between word vectors in a word2vec model etc.

## 5.3 Results

### *5.3.1 Applications - Genetics*

We first illustrate the performance of `CorShrink` on a data matrix with missing observations. The Genotype Tissue Expression (GTEx) Project [75] collected gene expression data from ≈ 540 subjects spanning across 51 different tissues and 2 cell lines. Different subjects, however, contributed different number of tissues leading to a large number of missing observations in the subject by tissue expression data matrix for each gene and hence differences in *matched samples* between tissues for computing the tissue-tissue correlation.

Figure 5.1 shows the results from `CorShrink` fit on the subject by tissue log CPM expression for *PLIN1* (*ENSG00000166819*) gene. Figure 5.1 (a) shows an image plot for the pairwise sample correlation of expression data between tissues, while 5.1 (c) displays the corresponding CorShrink fit using a mixture of half-uniform prior in Equation 5.7 - a choice that accounts for the fact that tissues in general are weakly positively correlated. The CorShrink estimated plot is visually more parsimonious and arguably more easily interpretable. The empirically fitted prior demonstrated higher concentration around small positive values (see Figure5.1 (b)). Expectedly the tissue pairs with low numbers of matched samples (almost white colored points) undergo high shrinkage while those with larger number of matched samples remain largely unperturbed by CorShrink (see Figure5.1 (d)) . Supplementary Figure S23 shows, for the same data as Figure 5.1, the results from applying CorShrink using other prior models - mixture normal prior centered around 0, mixture normal prior centered around a non-zero mode estimated from the data and non-parametric prior as defined in **Methods**.

Figure 5.1 presents a *tissue-wide* version of CorShrink that tries to shrink the tissue-tissue correlations for *PLIN1* gene based on the expression data for that gene alone. One may however consider determining the amount of shrinkage by leveraging information across all genes and to address this, we define a new *genewide* version of CorShrink. Under this model, for each tissue pair, we feed into Equation 5.6, the vector of pairwise correlation in expression for this tissue pair *across all genes*, together with another equal-sized vector, each of whose elements equals the number of matched samples for the tissue pair. Supplementary Figure S24 shows the image plots for both the *tissuewide* and *genewide* versions of CorShrink. Both these methods produce results that are similar to each other, but are less visually cluttered than the sample correlation estimate.

One characteristic feature of the tissue-tissue correlation structure of the *PLIN1* gene is the high correlation of expression in brain tissues. We observe this pattern in many genes, but there are exceptions. Supplementary Figure S25 presents the results of tissue-wide and gene-wide versions of CorShrink along with the pairwise sample correlation matrix for three different genes - *HBB*, *MTURN* and *VSIR*. *VSIR* correlation profile is similar to *PLIN1*, *HBB* gene exhibits high correlation in expression nearly across all tissues, while the *MTURN* gene exhibits low correlation in expression across all tissues.

One common way of dealing with missing data in statistics is to impute them. Factor analysis methods - such as SoftImpute [49, 80] and FLASH [133] attempt to fill in the missing values based on a lower dimensional representation of the data matrix estimated over the recorded observations. Supplementary Figure S26 presents, for the same *PLIN1* gene expression data, a comparison of the pairwise tissue-tissue correlation matrix (Fig 2a) and the corresponding tissue-wide `CorShrink` matrix (Fig 2b) with correlation matrix computed after imputing the missing observations in the data matrix using SoftImpute (Fig 2c) and FLASH (Fig 2d). The correlation matrices for both the SoftImpute and FLASH imputed data show an overall upward bias in their values - possibly driven by the large scale missing observations in the data.

### *5.3.2 Simulation studies*

Accounting for missing observations in the data matrix in an adaptive way for estimating the correlation matrix is the primary motivation behind using `CorShrink`. However, `CorShrink` is a very competent correlation shrinkage method even when there are no missing observations in the data matrix. Of particular interest are settings with small n (number of samples) and large p (number of features). Under various choices of $(n, p)$, with the ratio $n/p$ varying from 0.1 to 10, we performed simulation experiments to compare the performance of `CorShrink` against other popular correlation shrinkage methods - *GLASSO* at different tuning parameters [40, 81, 136], soft thresholding estimator *PDSCE* [101] and *corpcor* [106, 107].

We considered three types of underlying correlation structure to simulate from - a Hub correlation matrix (sparse correlation, sparse precision), a Toeplitz correlation matrix (sparse correlation, non-sparse precision) and a banded precision matrix (non-sparse correlation, sparse precision). See Supplementary Figure S27 for a demonstration of these correlation and precision matrix structures.

We generated multivariate normally distributed data with 0 mean and correlation structure determined as above. The number of features were fixed at $p = 100$ and four different values of $n$ were considered, $n = 10, 50, 100, 1000$. Figure 5.2 presents the box plot of the Correlation Matrix Distance (CMD) [50] between the population correlation matrix and the estimated matrices obtained using `CorShrink` (with normal mixture prior centered around 0), *GLASSO* at different tuning parameters, *PDSCE* and *corpcor* methods, together with the sample correlation matrix. Compared to the other approaches, the estimated matrix using `CorShrink` was observed to be closer (in CMD) to its population counterpart for the Hub and Toeplitz correlation models. For the sparse banded precision matrix case, GLASSO estimator performs better than the other estimators in small n, large p settings. Supplementary Figure S28 presents results for the same analysis, but using Frobenius distance metric instead of the CMD distance.

Figure 5.3 presents a second validation of performance of different estimators by comparing their trends in sorted eigenvalues of the estimated correlation matrices with that of the population correlation matrix. For the Toeplitz and the Hub correlation models, the `CorShrink` estimator appears to follow the population eigenvalues more closely than other methods for each choice of $n$ and $p$. The results of these simulation studies (Figures 5.2 and 5.3) seem to suggest that for small n, large p settings, with structured population correlation matrix, `CorShrink` outperforms the other methods, but for sparse precision models, *GLASSO* is the preferred choice. This is along expected lines, because`CorShrink` is more a correlation shrinkage model, whereas *GLASSO* is specifically designed for sparse representation of the precision matrix.

### *5.3.3 Applications - Natural Language Processing*

As discussed in the Methods section, `CorShrink` can be flexibly applied to other correlation-like quantities. One such application is in shrinking cosine similarities between vector representations of words, obtained from a *word2vec* [83] or *GLOVE* [87] model. The aim here is to generate more robust estimates of the cosine similarities between words, less affected by author, context or event specific biases.

As a case study, we considered text data from the monthly issues of the *Ebony* magazine in 1968. 1968 marked a turning point in American history with the assassination of Dr. Martin Luther King and the subsequent end to the civil rights movement and the Ebony magazine issues provided a reflection of those times. We fitted *word2vec* model on the combined text data from these issues, obtained vector representations and computed cosine similarities between words based on the vector representations. These cosine similarities are treated as correlation like quantity in the `CorShrink` model.

Unlike correlations, there is no obvious way to compute standard errors of the Fisher Z-scores for these cosine similarities, which made us resort to re-sampling methods. We performed Bootstrapping [28, 31] on the 12 issues, fitted *word2vec* model on the pooled text from the re-sampled articles and computed cosine similarities of our chosen food-related words from their model vector representations. For each pair of words, we computed a re-sampling standard error of the Fisher Z-scores from 100 re-samples.

Two word sets we were interested in were {*martin, luther, king*} and {*civil, rights*}. For each word set of interest, we selected top 1000 words close in context to the words in these word sets based on cosine-similarities from the word2vec model fit and then combined these words. Next, `CorShrink` was applied to these word pairs with re-sampling standard errors computed as above. Figure 5.4 presents the original and CorShrink cosine similarity patterns along with the top 25 words contextually similar to our word sets of interest based on cosine similarities and `CorShrink` estimated similarities.

The word rankings after `CorShrink` adjustment seem to give higher preference to terms that are broadly contextually similar to the words in the word sets. For example words like *peace*, *civil* and *rights* show up among the top 25 words close to the word set *martin, luther, king* after `CorShrink` adjustment but do not show up before the adjustment. Also, a term like *apostle* which seems strongly connected to this word set before CorShrink (ranked 4) disappears from top 25 words list after adjustment. Upon investigation, we found this word to be used primarily in the context of eulogizing Dr. King following his death in the May 1968 edition issue, and re-sampling on the articles managed to remove this bias. Similarly, the top 25 words contextually close to *civil, rights* before `CorShrink` adjustment seem to consist of names like *andrew* and *randolph*, which are apparently names used in the context of civil rights in specific issues and also, surprisingly a word like *cowboys* which is quite distant in context from the word set of interest. The `CorShrink` adjustment cleans out all these words from the top 25 words list and the terms included instead are again broadly related to *civil* and *rights* - like *militant*, *war*, *freedom* etc.

## 5.4 Discussion

`CorShrink` is an extension of the adaptive shrinkage (*ash*) framework by [119] to the setting of correlation matrix shrinkage. Unlike other correlation matrix estimation approaches (*corpcor*, *GLASSO*), `CorShrink` can adjust the degree of shrinkage based on the missing observations in the data (see Figure 5.1), and even with no missing data, outperforms the other methods when the underlying population correlation matrix is well-structured (see Figure 5.2) Also, while other methods take only a data matrix as input, `CorShrink` has the flexibility to take as input either a vector/matrix of correlations along with the information of matched samples, which is what extends the flexibility of this approach to shrinking cosine similarities between word pairs from text data. In terms of speed, `CorShrink` is comparable to *corpcor*, *PDSCE* [106] and *glasso* [40], but unlike *glasso* and *PDSCE*, does not require extensive training of the tuning parameter using cross-validation. As argued in the Results, the current implementation of `CorShrink` is not ideal for estimating inverse correlations, and our future works would be directed towards improving the efficiency of `CorShrink` in estimating precision matrices. Also, when the population correlation matrix is a noisy version of a lower dimensional structured matrix, `CorShrink` is not well suited to recover the lower dimensional structure. Future attempts would focus on combining `CorShrink` with factor analysis type approaches to recover this lower dimensional structure. `CorShrink` is currently available as a R package on Github `https://github.com/kkdey/CorShrink` and the codes for the analysis presented in this paper are available at `https://kkdey.github.io/CorShrink-pages/`.

## 5.5 Author contributions

Dey, KK and Stephens, M designed the method. Dey, KK implemented the method. Dey, KK ran the experiments. Dey, KK produced the figures. Dey, KK and Stephens, M wrote the paper.

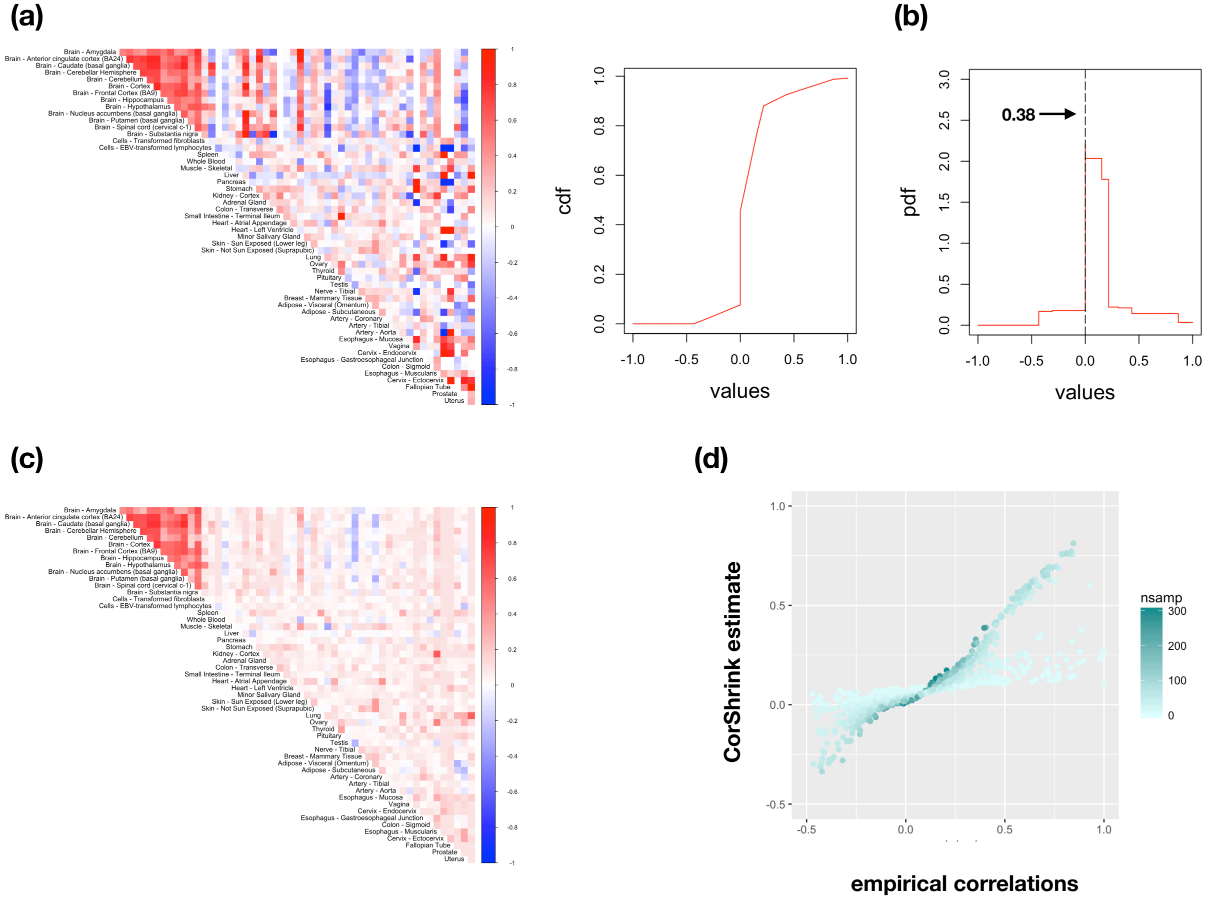


Figure 5.1: (a) The image plot of the pairwise correlation matrix between tissue pairs for the log CPM expression data of the *PLIN1* gene. (b) The probability and cumulative density function plots for the empirically estimated mixture of half-uniform prior used to shrink the correlations in CorShrink. The black dot represents the prior probability mass of observing 0 correlation. (c) The image plot the estimated correlation matrix due to CorShrink. The representation is visually more parsimonious and arguably more interpretable than (a). (d) plots the pairwise sample correlation values against the CorShrink fitted estimates for each tissue pair in a scatter plot and colors each point based on the number of matched samples for the corresponding tissue pair. Expectedly the pairs with low numbers of matched samples (light colored points) undergo high shrinkage while those with larger number of matched samples remain largely unperturbed by CorShrink.

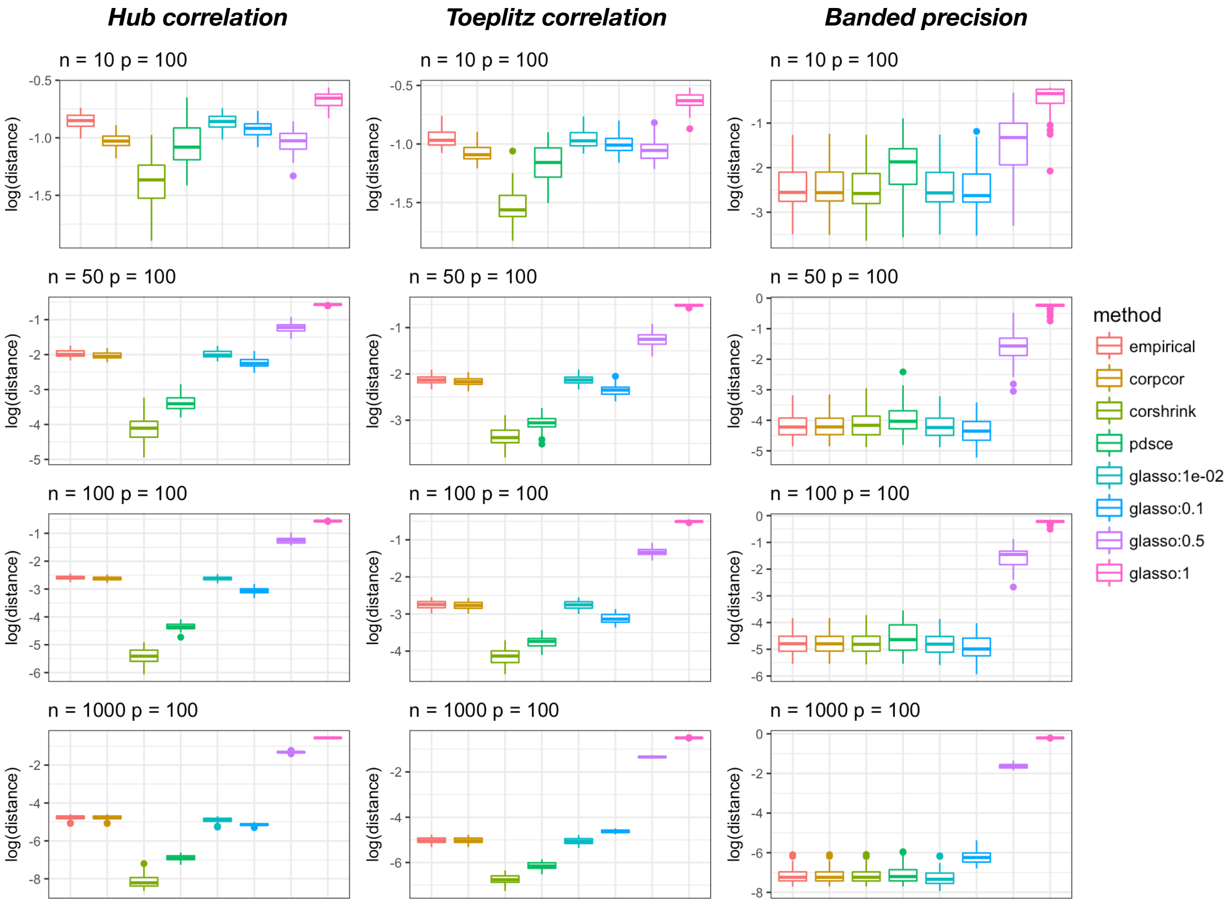


Figure 5.2: Box plot of the Correlation Matrix Distance (CMD) [50] between population correlation matrix and the estimated matrix from different methods - *corpcor*[106, 107], `CorShrink`, *PDSCE* [101], GLASSO [40] at different tuning parameters and the empirical pairwise correlation matrix, for different structural assumptions on the underlying population correlation - Hub structure, Toeplitz structure and a banded precision matrix, see Supplementary Figure S27. `CorShrink` outperforms the other methods for the structured/sparse covariance models (Hub and Toeplitz), with *PDSCE* being closest to `CorShrink` in performance.

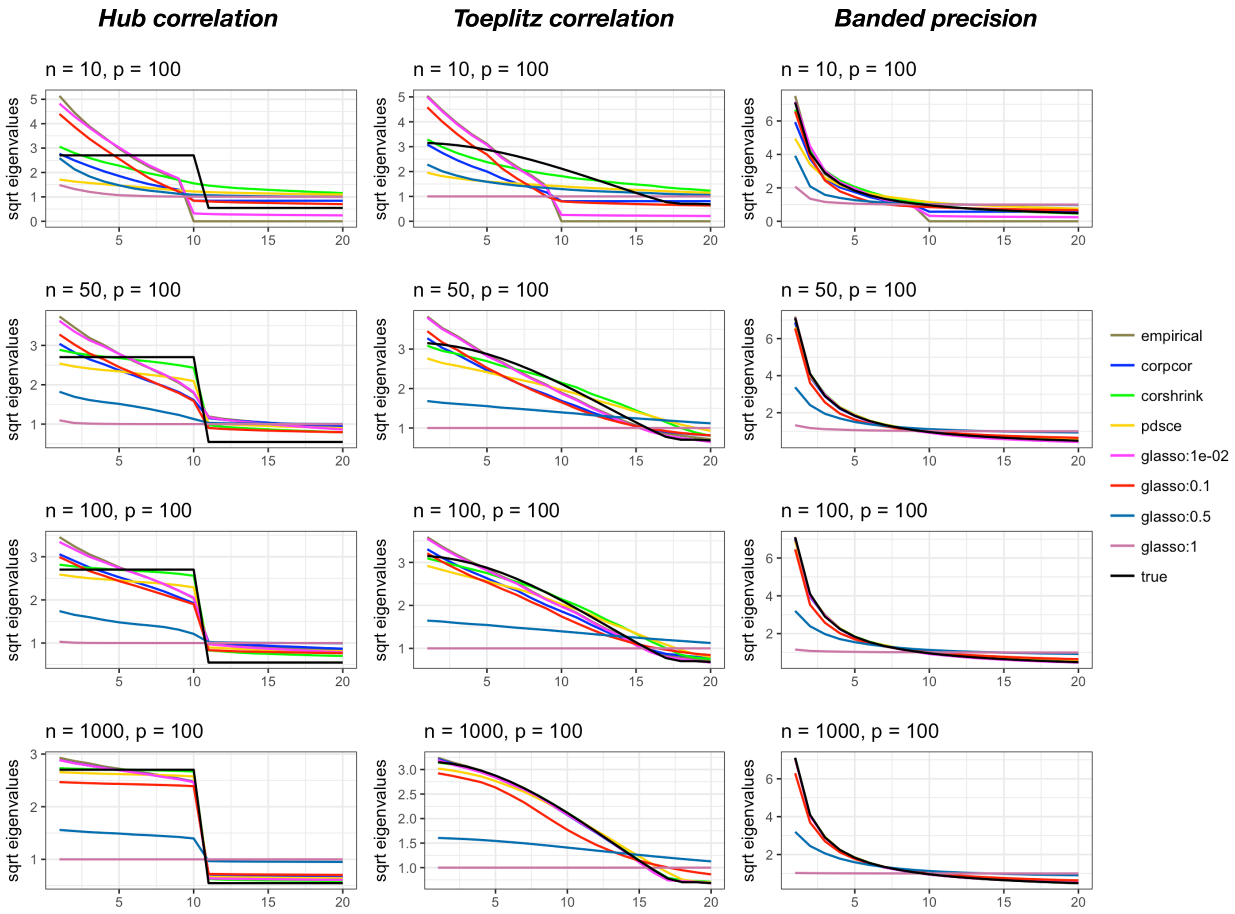
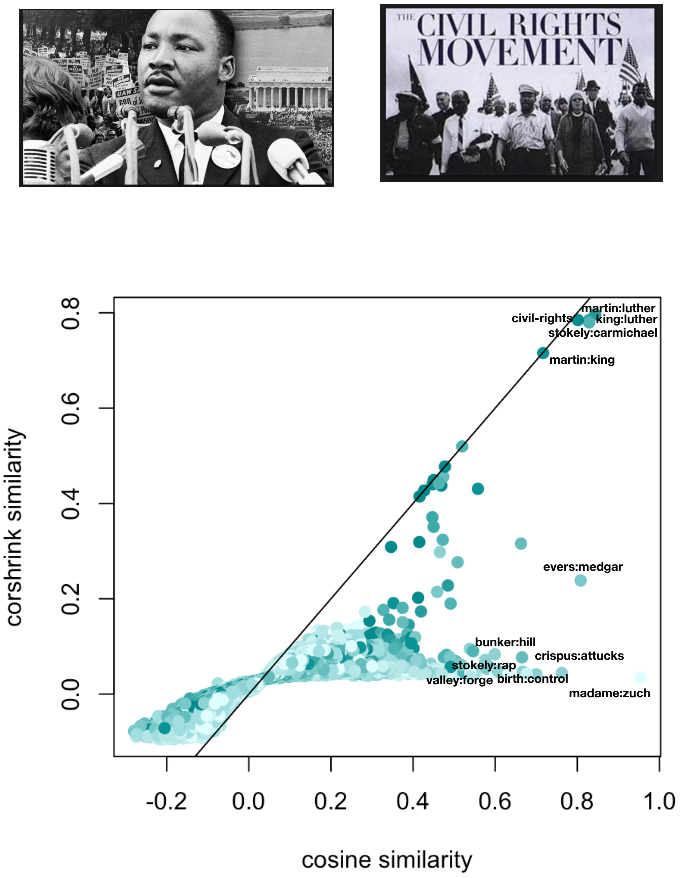


Figure 5.3: Plots of sorted square-root eigenvalue trends of the population correlation matrix against those of estimated correlation matrices using different methods - *corpcor* [106, 107], `CorShrink`, *PDSCE* [101], GLASSO [40] at different tuning parameters and the empirical pairwise correlation matrix, for different structural assumptions on the underlying population correlation - hub structure, Toeplitz structure and a banded precision matrix structure, see Supplementary Figure S27. The trend of sorted eigenvalues for `CorShrink` follow that of the original matrix closely for the Hub and the Toeplitz models.



Words close in context to (martin, luther, king)

***before CorShrink***

```
   luther       martin         king          rev      apostle        floyd    requested
0.8908471    0.8535049    0.8485832    0.4063000    0.3981613    0.3509520    0.3498945
    fiery  assassinated      murder      francis      funeral       forres         late
0.3463924    0.3427114    0.3375902    0.3301637    0.3293240    0.3269475    0.3180746
    kings         joan       prince     caucuses    abernathy    preaching      marched
0.3070072    0.3051331    0.3048364    0.2954050    0.2925116    0.2891073    0.2868515
  prophet      kennedy assassination      seeds
0.2835925    0.2812839    0.2809260    0.2785261
```

***after CorShrink***

```
    luther       martin         king          rev      jackson       leader assassination
0.86047946   0.83742984   0.83344445   0.23006744   0.10901378   0.09842106   0.09665314
     james        kings        ralph   montgomery   courtship         late   birmingham
0.09482433   0.09374696   0.09264236   0.09258621   0.09234065   0.09025221   0.09022915
    murder        civil        peace     actively       wright     personal       rights
0.08954815   0.08924835   0.08922241   0.08921495   0.08874579   0.08867831   0.08789465
     naacp          god        voice      marched
0.08683455   0.08661692   0.08588993   0.08563276
```

Words close in context to (civil, rights)

***before CorShrink***

```
         civil          rights        movement     legislation       disorders           bills
     0.9007841       0.9007841       0.4405455       0.3730043       0.3550964       0.3421801
   enforcement       movements    strengthened     involvement           equal      protection
     0.3402390       0.3323286       0.3269478       0.3155524       0.3120888       0.3120654
reconstruction        belonged         cowboys          andrew        randolph      amendments
     0.3117899       0.3022773       0.2964739       0.2961406       0.2911620       0.2899793
       murders      commission      nonviolent recommendations          brutal            sncc
     0.2860052       0.2846442       0.2844914       0.2823338       0.2760409       0.2733980
       paradox
     0.2722072
```

***after CorShrink***

```
     civil       rights     movement        equal     militant   commission          war
0.90078411   0.90078411   0.28434679   0.11712699   0.11265973   0.11157274   0.10768549
   freedom    committee  involvement       luther     equality   principles       became
0.10677891   0.10644511   0.10586616   0.10202443   0.10028706   0.09897363   0.09890537
     human complications     georgia constitution  nonviolence  legislation       voting
0.09883332   0.09716990   0.09696635   0.09610843   0.09607732   0.09593372   0.09576468
   workers  integration    political          law
0.09572006   0.09560082   0.09551930   0.09485102
```

Figure 5.4: We extracted the top 1000 contextually close words to each of the two word sets of our interest - {*martin, luther, king*} and {*civil, rights*} based on *word2vec* analysis of the monthly issues of the Ebony magazine in 1968 and combined these two sets of words. For each pair of words in the combined word set, we plotted the cosine similarities before and after the `CorShrink` adjustment, colored by how many times they occurred in the texts. We report the top 25 words contextually close to the word sets of interest before and after the `CorShrink` adjustment. The `CorShrink` adjustment seems to remove terms that are specific to a few texts or specific to certain events and instead incorporate more broadly similar words to our word sets of interest in the top 25 lists. Examples are discussed in depth in the Results section.

# Appendix A

## Supplementary Figures

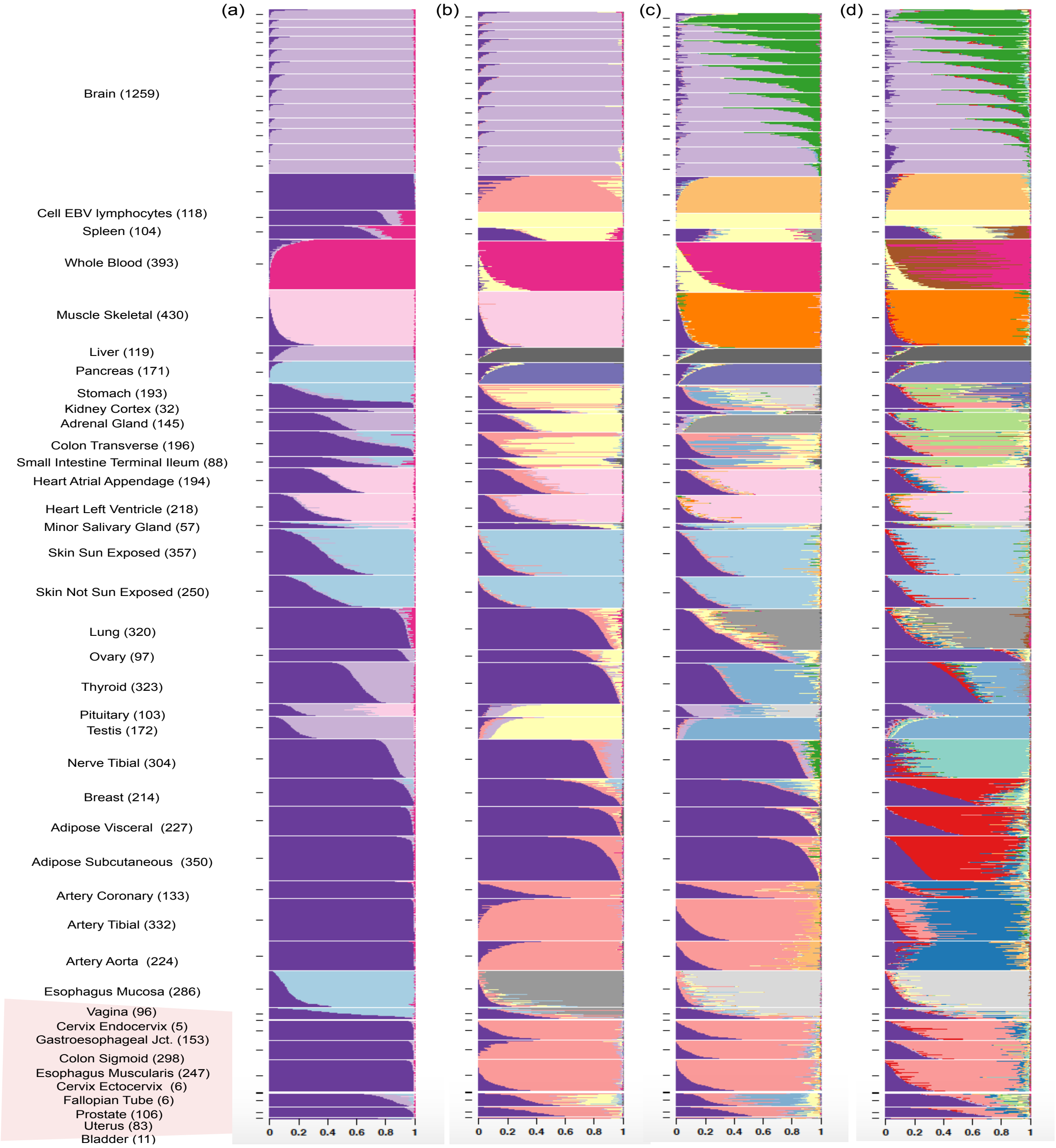


Figure S1: **Structure plot of GTEx V6 tissue samples for (a) $K = 5$, (b) $K = 10$, (c) $K = 15$, (d) $K = 20$.** Some tissues form a separate cluster from the other tissues from $K = 5$ onwards (for example: Whole Blood, Skin), whereas some tissue only form a distinctive subgroup at $K = 20$ (for example: Arteries).

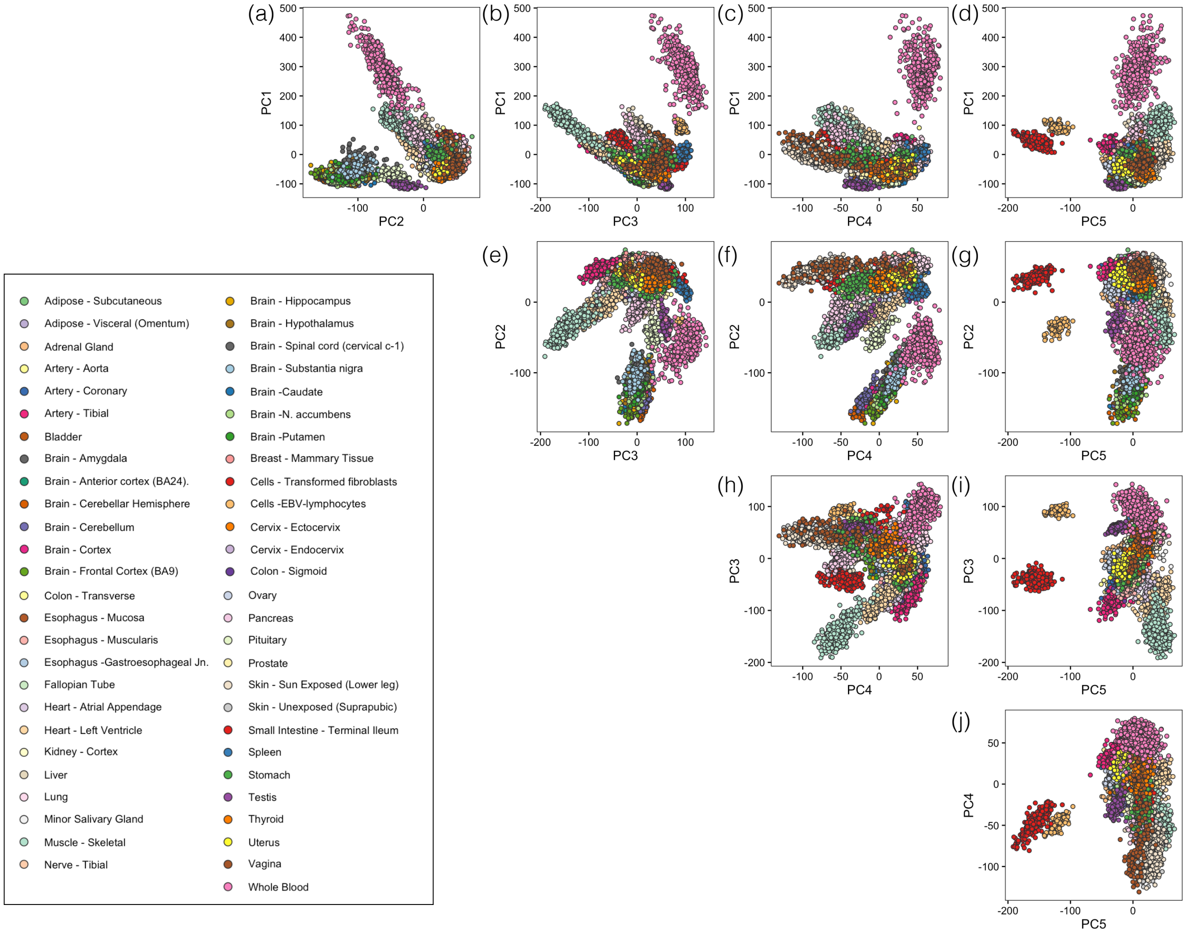
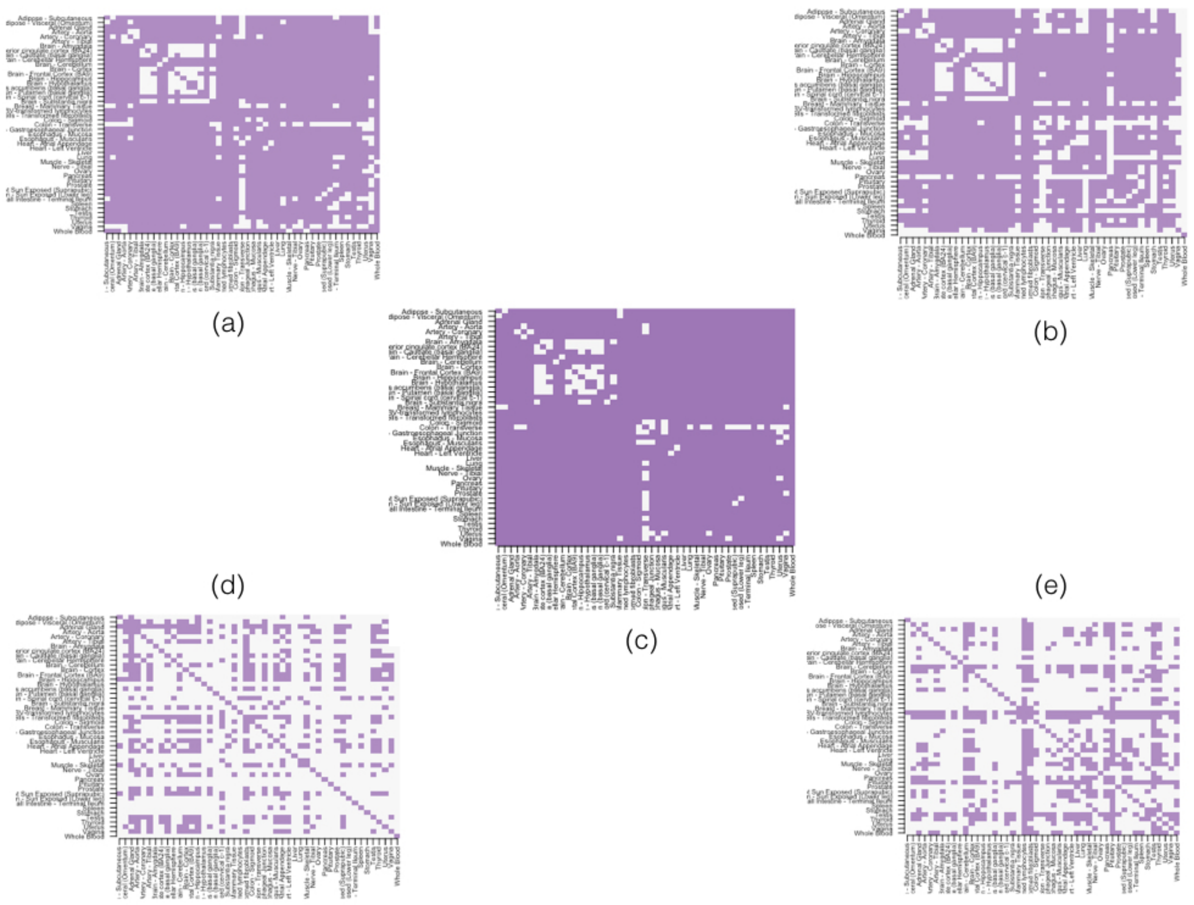


Figure S2: **Top five principal components (PC) for GTEx V6 tissue samples.** Scatter plot representation of the top five PCs of the GTEx tissue samples. Data was transformed to log2 counts per million (CPM).



(a)

(b)

(c)

(d)

(e)

Figure S3: **Comparison between GoM model and hierarchical clustering under different scenarios of data transformation.** We used GTEx V6 data for model performance comparisons. Specifically, for every pair of the 53 tissues, we assessed the ability of the methods to separate samples according to their tissue of origin. The subplots of heatmaps show the results of evaluation under different scenarios. Filled squares in the heatmap indicate successful separation of the samples in corresponding tissue pair comparison. (a) Hierarchical clustering on log2 counts per million (CPM) transformed data using Euclidean distance. (b) Hierarchical clustering on the standardized log2-CPM transformed data (transformed values for each gene was mean and scale transformed) using the Euclidean distance. (c) GoM model of $K = 2$ applied to counts. (d) Hierarchical clustering on counts data with the assumption that, for each gene the sample read count $c_{ng}$ has a variance $\bar{c}_g + 1$. (e) Hierarchical clustering applied to adjusted count data. Each gene has a mean expression value of 0 and variance of 1. The GoM model with $K = 2$ is able to separate samples of different tissue of origin, better than hierarchical cluster methods.

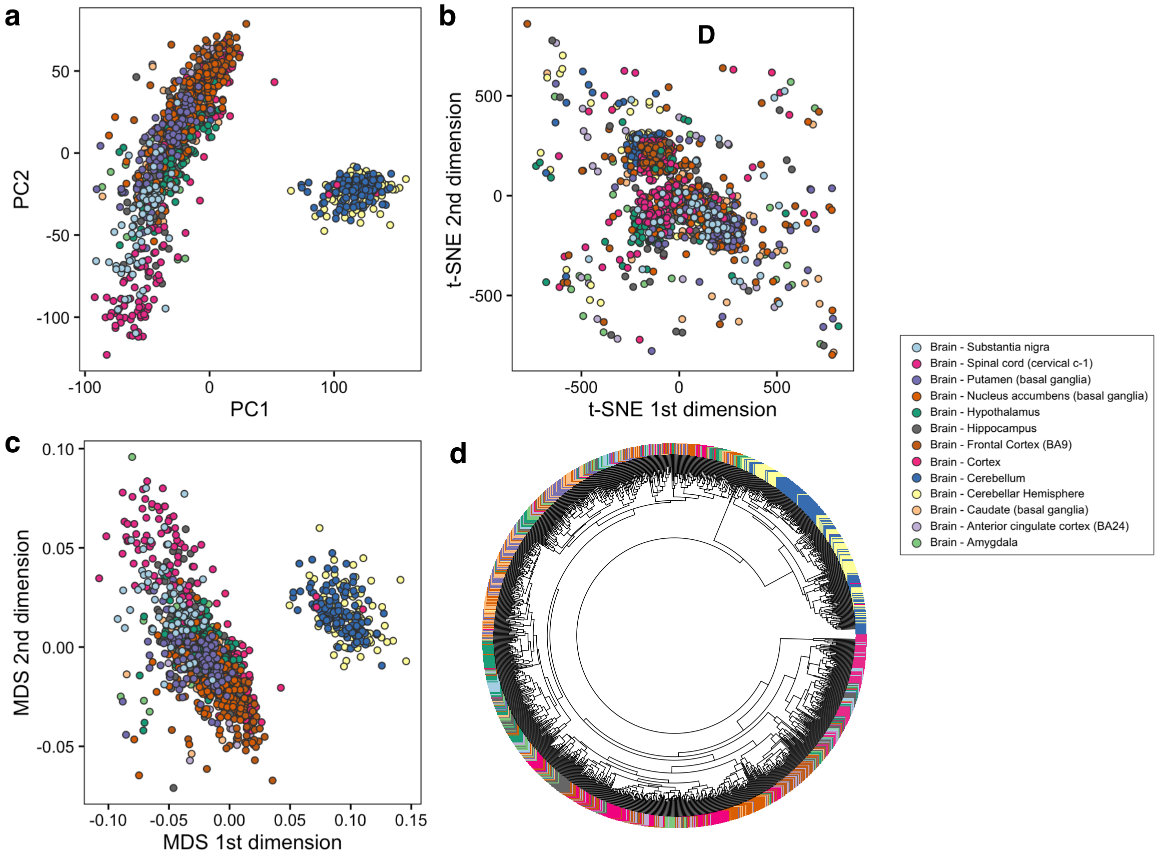


Figure S4: **GTEx brain PCA, t-SNE and MDS.**

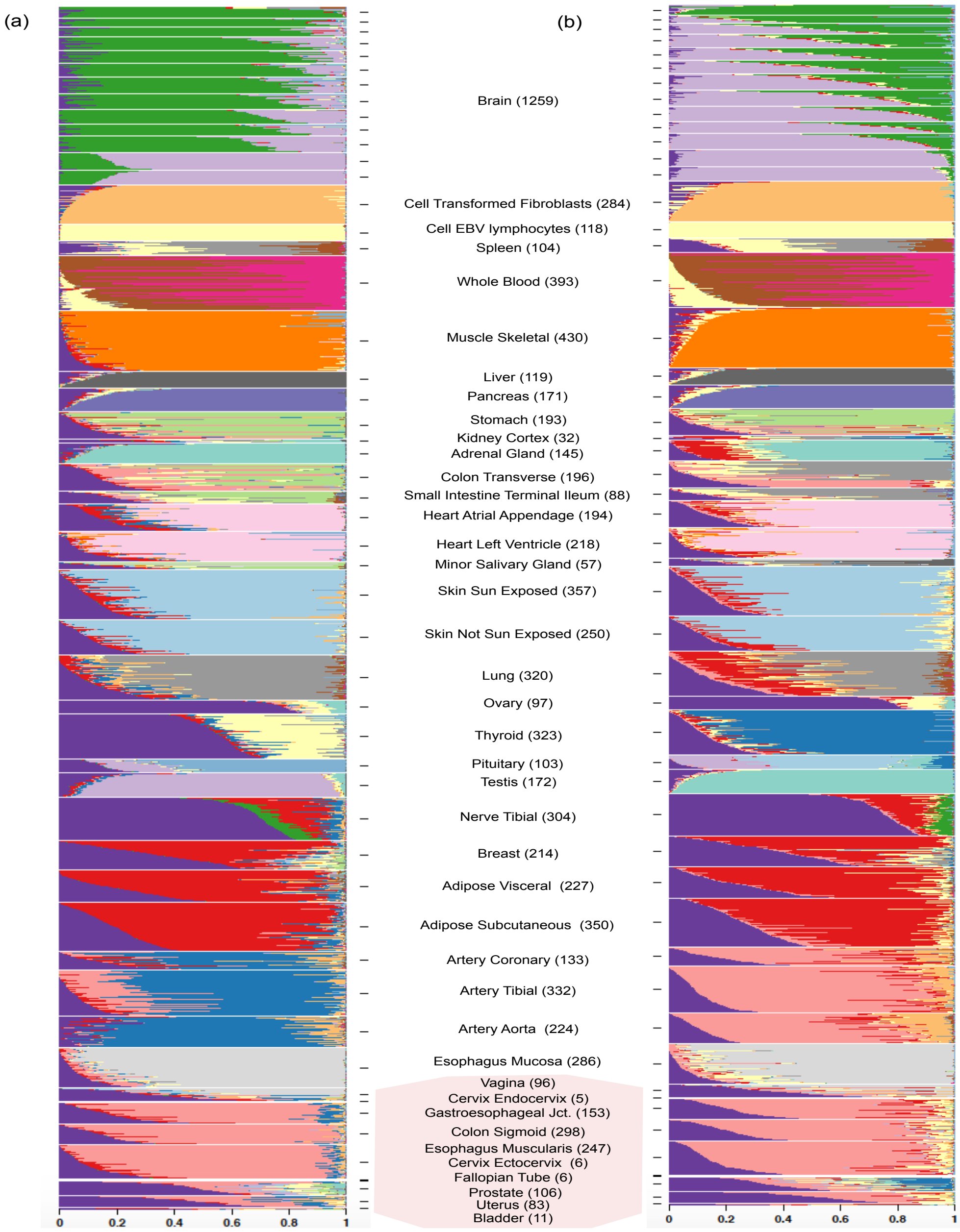
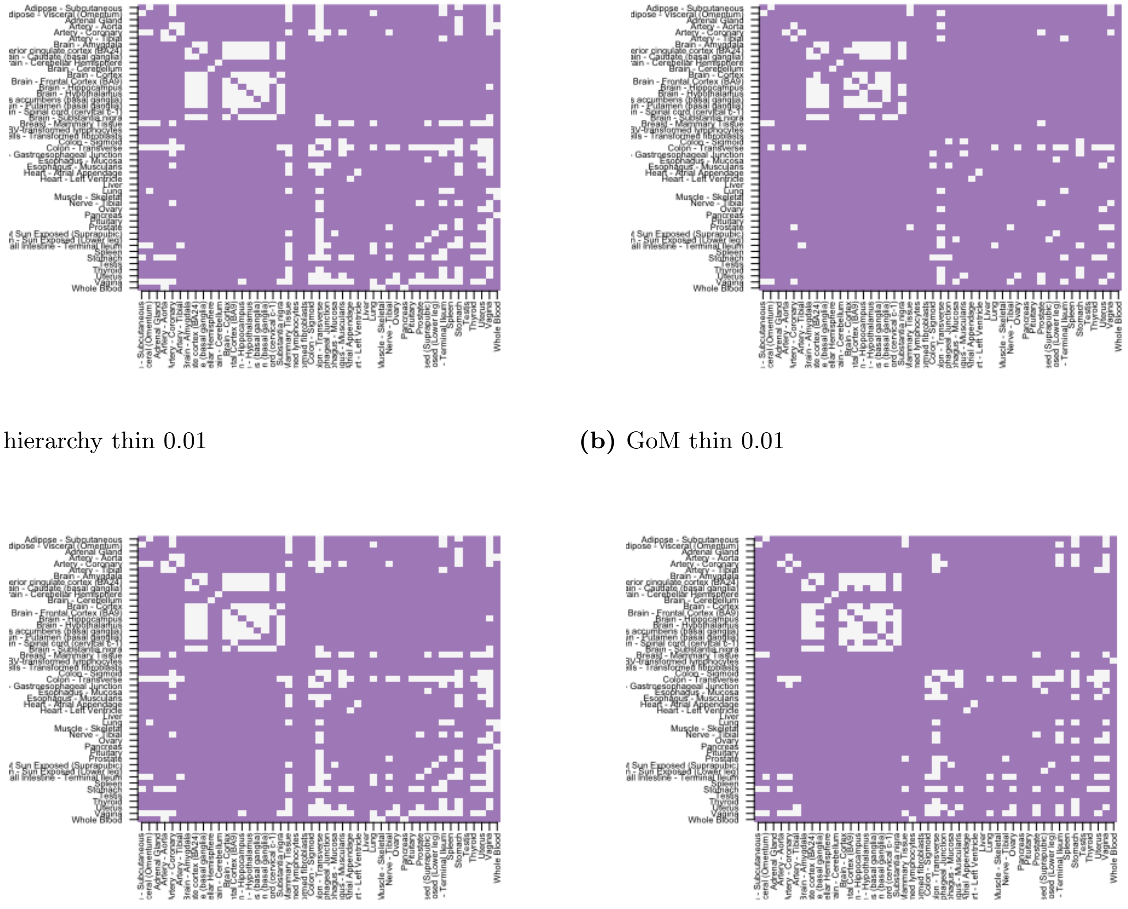


Figure S5: **GTEx brain PCA, t-SNE and MDS.**



**(a)** hierarchy thin 0.01

**(b)** GoM thin 0.01

**(c)** hierarchy 0.001

**(d)** GoM thin 0.001

Figure S6: **A comparison of accuracy of hierarchical clustering vs GoM on thinned GTEx data, with thinning parameters of $p_{thin} = 0.01$ and $p_{thin} = 0.001$.** For each pair of tissue samples from the GTEx V6 data we assessed whether or not each clustering method (with $K = 2$ clusters) separated the samples according to their tissue of origin, with successful separation indicated by a filled square. Thinning deteriorates accuracy compared with the unthinned data (Fig 2), but even then the model-based method remains more successful than the hierarchical clustering in separating the samples by tissue or origin

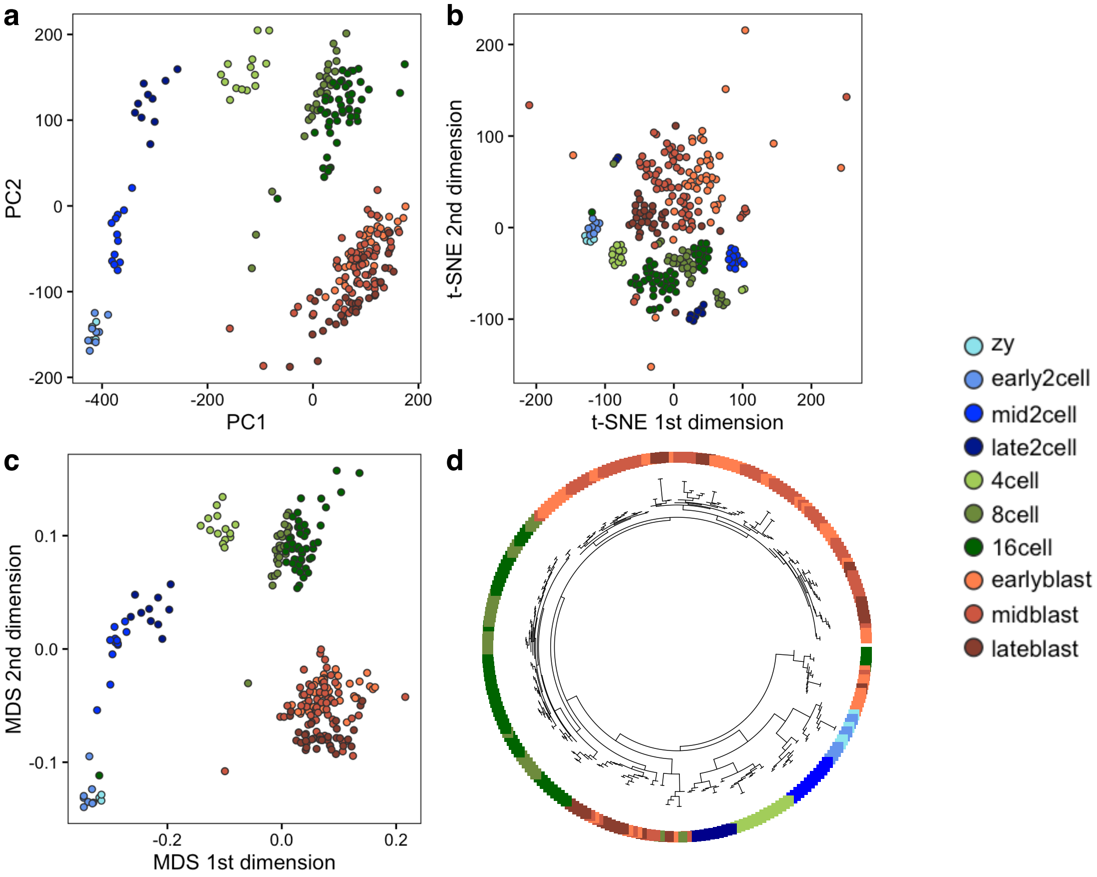
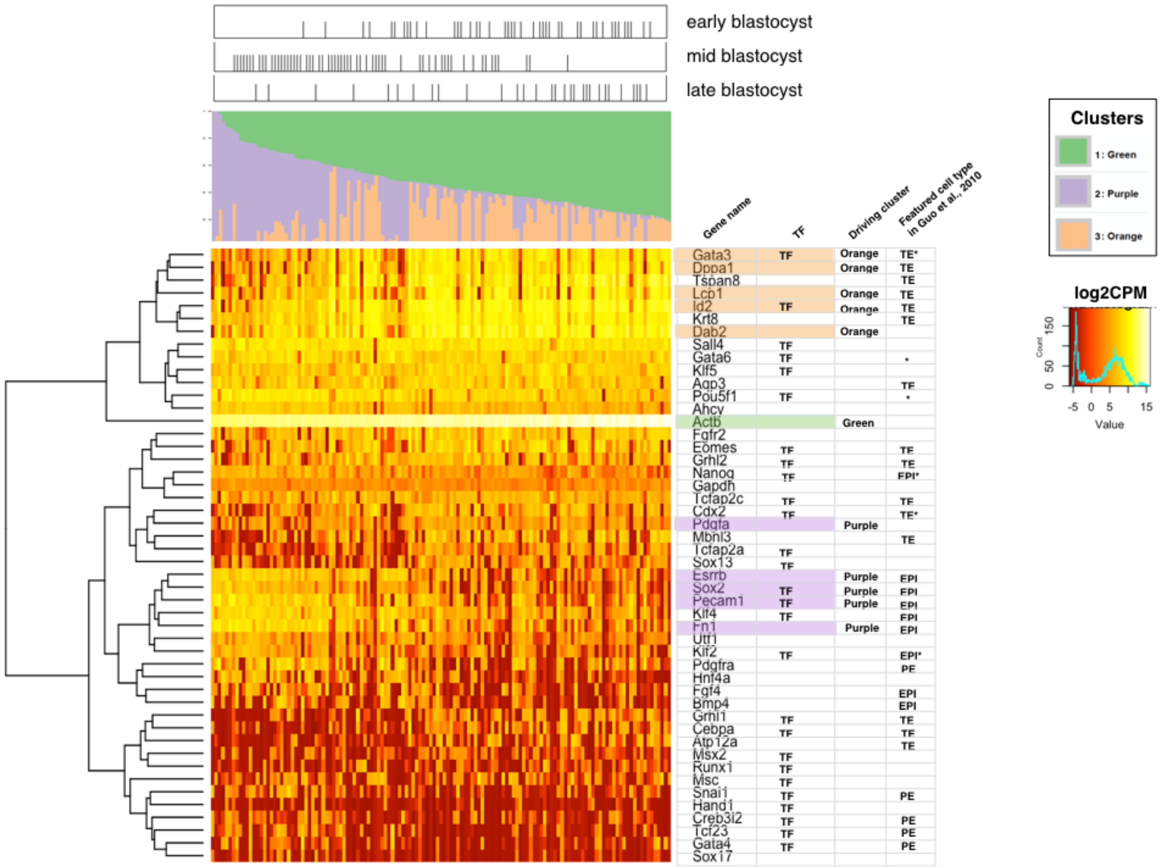


Figure S7: **Deng et al (2014) PCA, tSNE, MDS and dendrogram plots for hierarchical clustering.**



| Gene name | TF | Driving cluster | Featured cell type in Guo et al., 2010 |
|---|---|---|---|
| Gata3 | TF | Orange | TE* |
| Dppa1 | | Orange | TE |
| Tspan8 | | | TE |
| Lcp1 | | Orange | TE |
| Id2 | TF | Orange | TE |
| Krt8 | | | TE |
| Dab2 | | Orange | |
| Sall4 | TF | | |
| Gata6 | TF | | * |
| Klf5 | TF | | |
| Aqp3 | | | TF |
| Pou5f1 | TF | | * |
| Ahcy | | | |
| Actb | | Green | |
| Fgfr2 | | | |
| Eomes | TF | | TF |
| Grhl2 | TF | | TE |
| Nanog | TF | | EPI* |
| Gapdh | | | |
| Tcfap2c | TF | | TF |
| Cdx2 | TF | | TE* |
| Pdgfa | | Purple | |
| Mbnl3 | | | TE |
| Tcfap2a | TF | | |
| Sox13 | TF | | |
| Esrrb | | Purple | EPI |
| Sox2 | TF | Purple | EPI |
| Pecam1 | TF | Purple | EPI |
| Klf4 | TF | | EPI |
| Fn1 | | Purple | EPI |
| Utf1 | | | |
| Klf2 | TF | | EPI* |
| Pdgfra | | | PE |
| Hnf4a | | | |
| Fgf4 | | | EPI |
| Bmp4 | | | EPI |
| Grhl1 | TF | | TE |
| Cebpa | TF | | TF |
| Atp12a | | | TE |
| Msx2 | TF | | |
| Runx1 | TF | | |
| Msc | TF | | |
| Snai1 | TF | | PE |
| Hand1 | TF | | |
| Creb3l2 | TF | | PE |
| Tcf23 | TF | | PE |
| Gata4 | TF | | PE |
| Sox17 | | | |

Figure S8: **Additional GoM analysis of Deng et al (2014) data including blastocyst samples and 48 blastocyst marker genes.** We considered 48 blastocyst marker genes (as chosen by Guo et al., 2010) and fitted GoM model with $K = 3$ to 133 blastocyst samples. In the Structure plot, blastocyst samples are arranged in order of estimated membership proportion in the Green cluster. The panel located above the Structure plot shows the corresponding pre-implantation stage from which blastocyst samples were collected. The heatmap located below the Structure plot represents expression levels of the 48 blastocyst marker genes (log2 CPM), and the corresponding dendrogram shows results of hierarchical clustering (complete linkage). The table on the right of the expression heatmap displays gene information, showing, from left to right, 1) whether or not the gene is a transcription factor, 2) the driving GoM cluster if the gene was among the top five driving genes, and 3) the featured cell type (TE: trophecoderm, EPI: epiblast, PE: primitive endoderm) that was found in Guo et al., 2010.

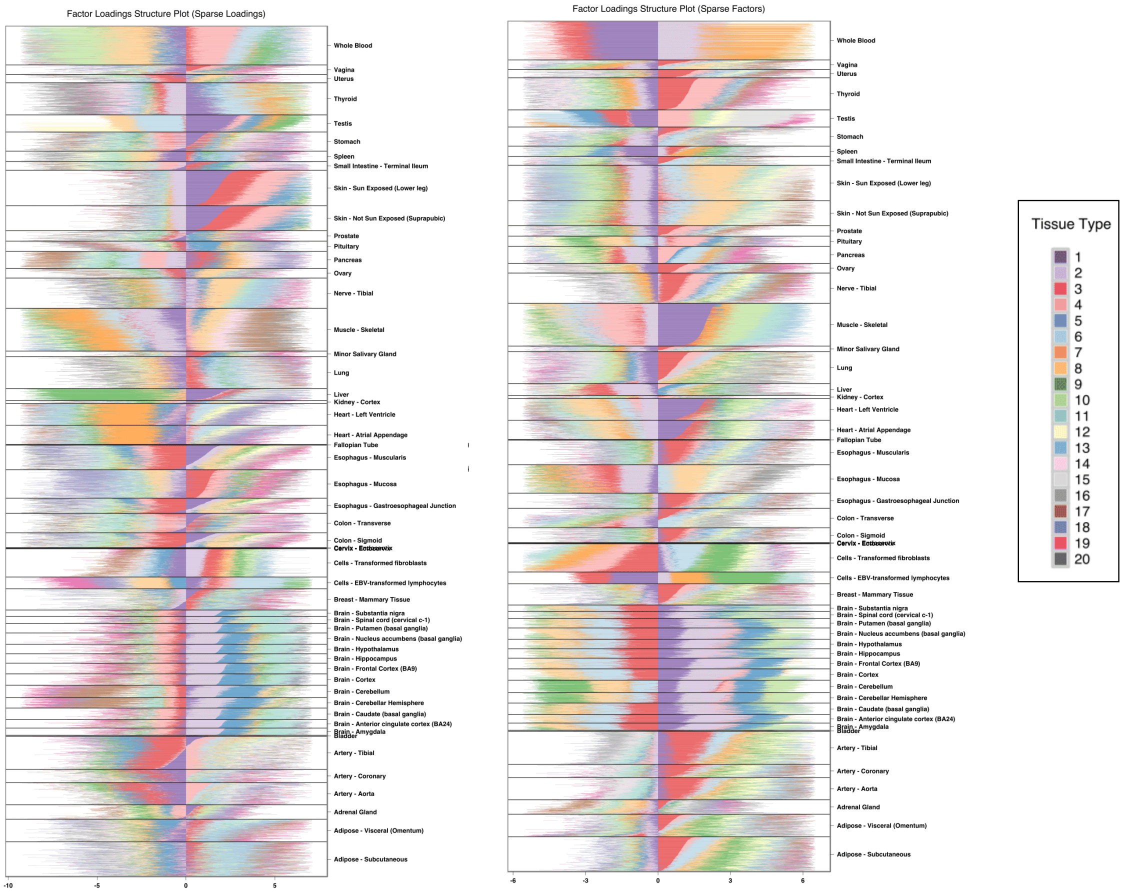


Figure S9: **Sparse Factor Analysis loadings visualization of GTEx V6 tissue samples.** The colors represent the 20 different factors. The factor loadings are presented in a stacked bar for each sample. We performed SFA under the scenarios of when the loadings are sparse (left panel) and when the factors are sparse (right panel).

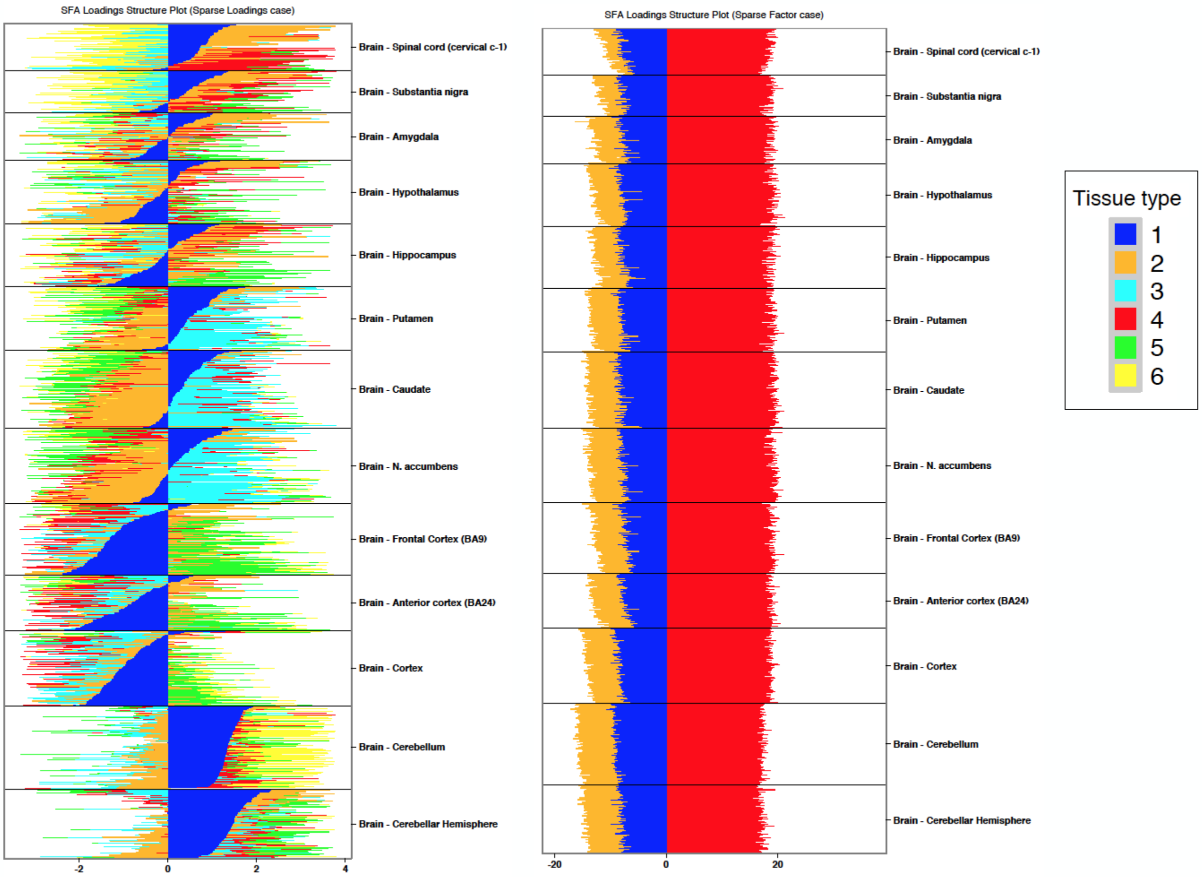


Figure S10: **Sparse Factor Analysis loadings visualization of GTEx brain tissue samples.** The colors represent the 6 different factors. The factor loadings are presented in a stacked bar for each sample. We performed SFA under the scenarios of when the loadings are sparse (left panel) and when the factors are sparse (right panel).

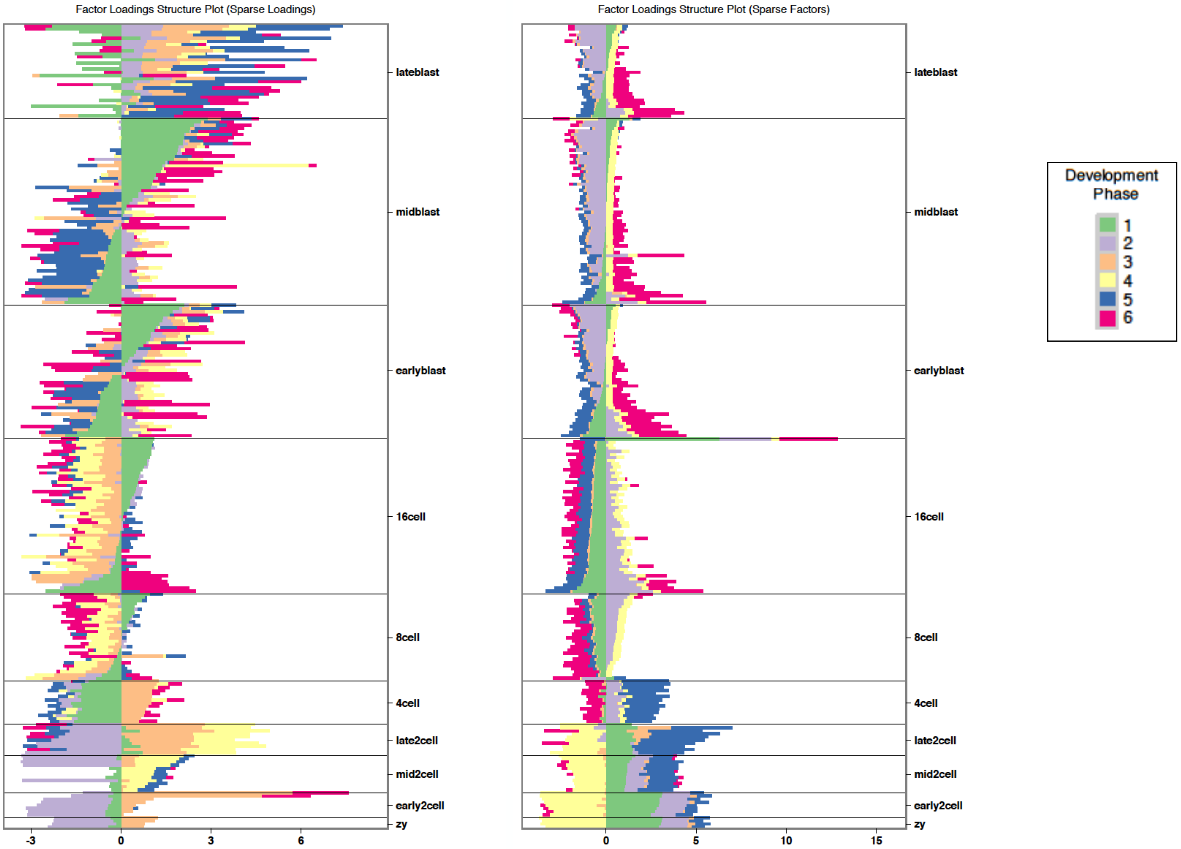


Figure S11: **Sparse Factor Analysis loadings visualization of mouse pre-implantation embryos from Deng et al., (2014). The colors represent the 6 different factors. The factor loadings are presented in a stacked bar for each sample. We performed SFA under the scenarios of when the loadings are sparse (left panel) and when the factors are sparse (right panel).**

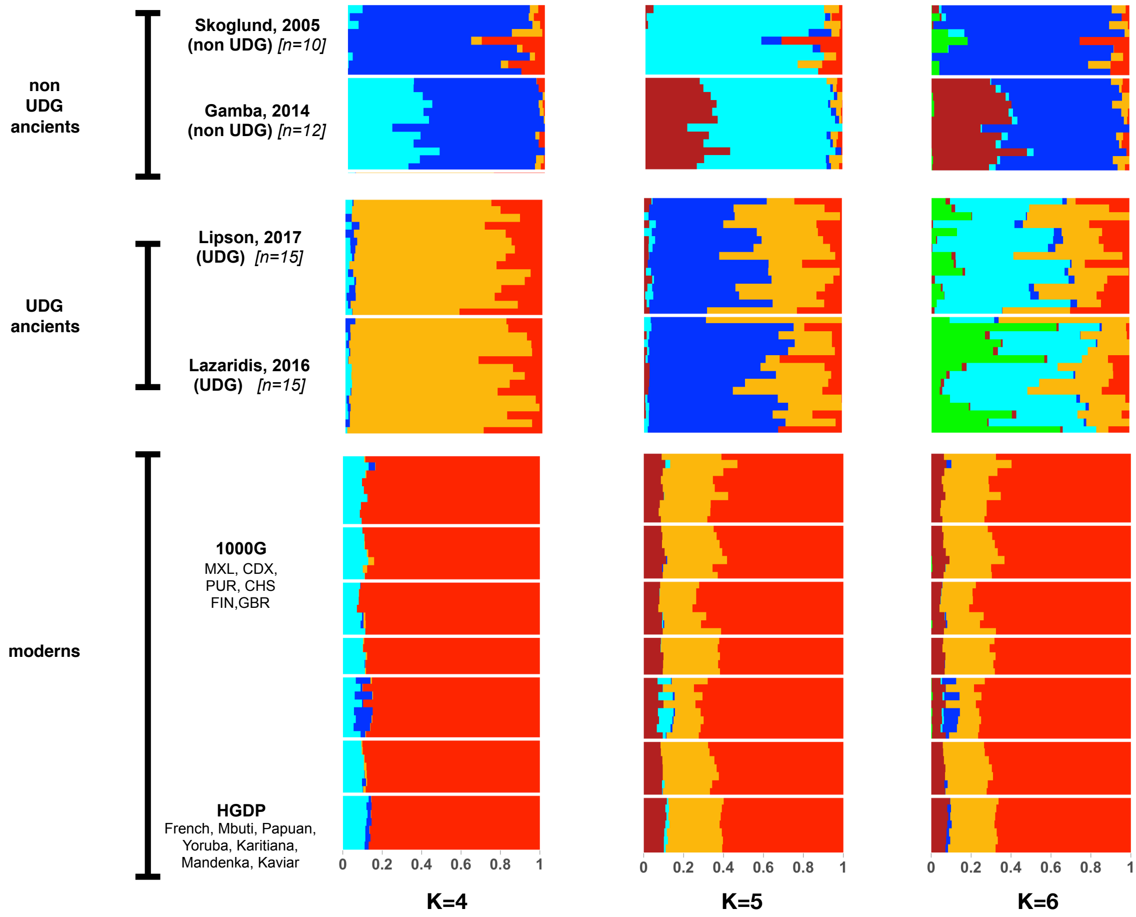


Figure S12: `aRchaic` grades of membership for the example in Fig 4.2 corresponding to 3 different values of $K$ ($K = 4, 5, 6$). Higher values of $K$ distinguish among the ancient studies, reflecting lab and study specific biases.

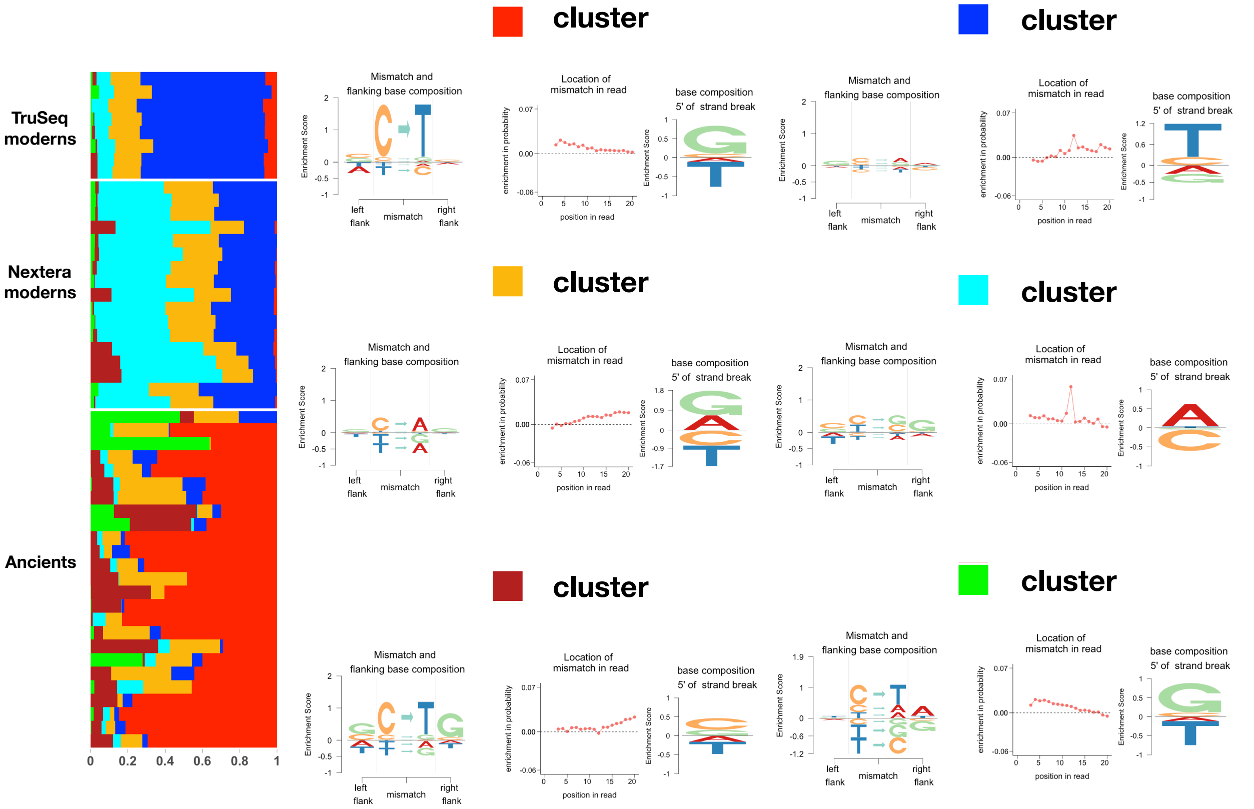


Figure S13: We apply `aRchaic` with $K = 6$ on the data from Fig 3.4. In addition to separating out the ancients from the moderns, `aRchaic` now distinguishes between moderns individuals based on library kit (Nextera vs Tru-seq.)

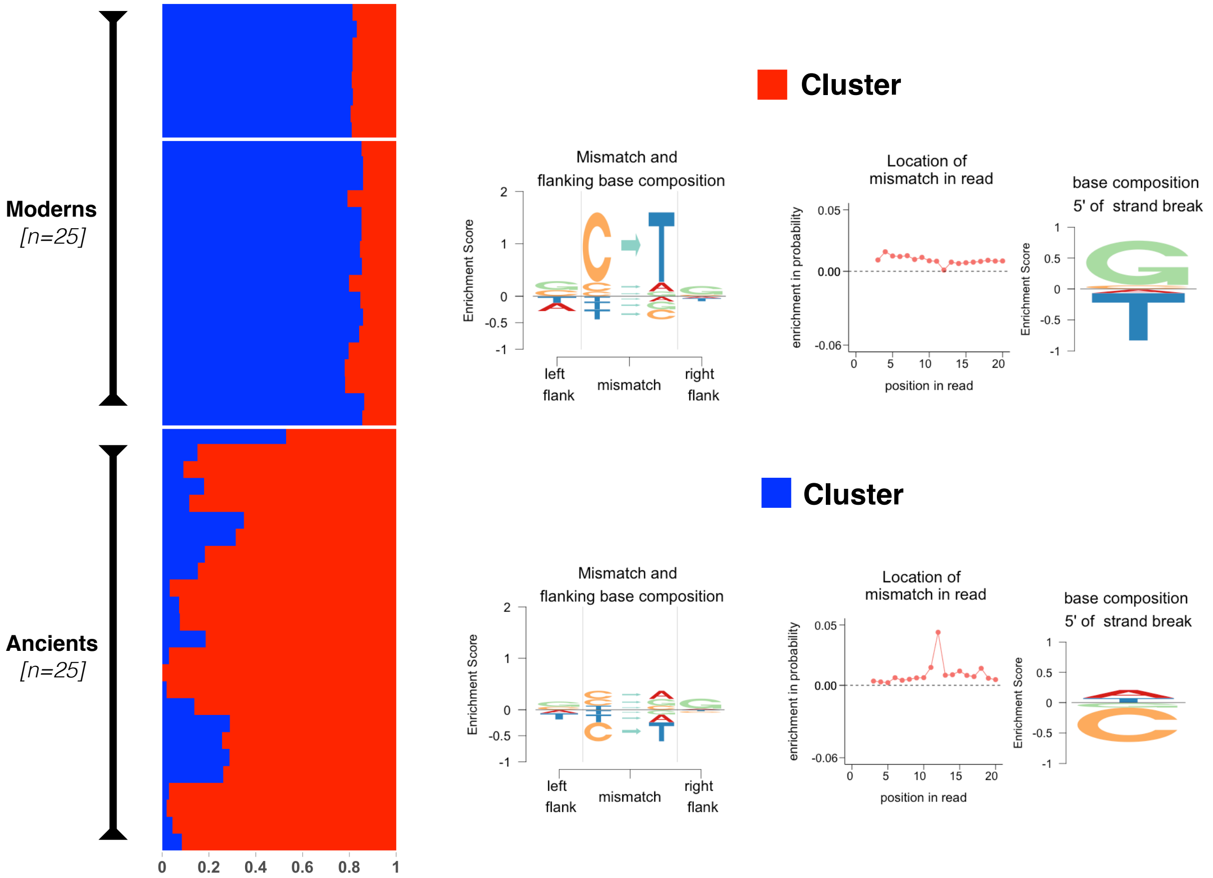


Figure S14: `aRchaic` plot for $K = 2$ on the combined data of 25 moderns and 25 ancients from [73]. `aRchaic` clearly distinguishes the moderns from the ancients. The ancients are primarily presented by the blue cluster. This cluster shows an enrichment of $C \rightarrow T$ mismatches and depletion of $T \rightarrow C$ mismatches with respect to modern background, as well as enrichment of G and depletion of T at the 5' strand break. The red cluster shows a blip at 12th position from the end of the read, the explanation for which is provided S15.

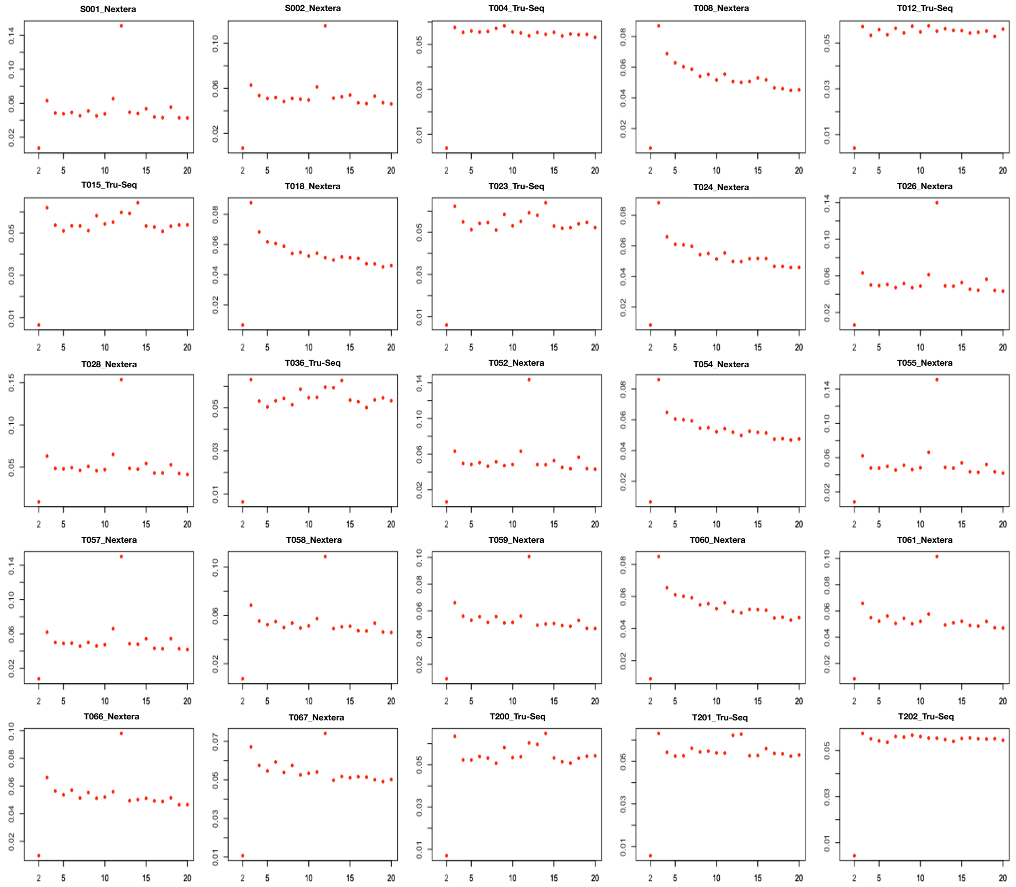


Figure S15: The frequency of all mismatch types plotted against the position of the read (from the 5' end) for each of the 25 moderns samples in [73]. Each sample was prepared by one of two library kits: Nextera and True-Seq. Most the samples prepared with the Nextera kit show a spike in frequency at the 12th position from the 5' end of the read

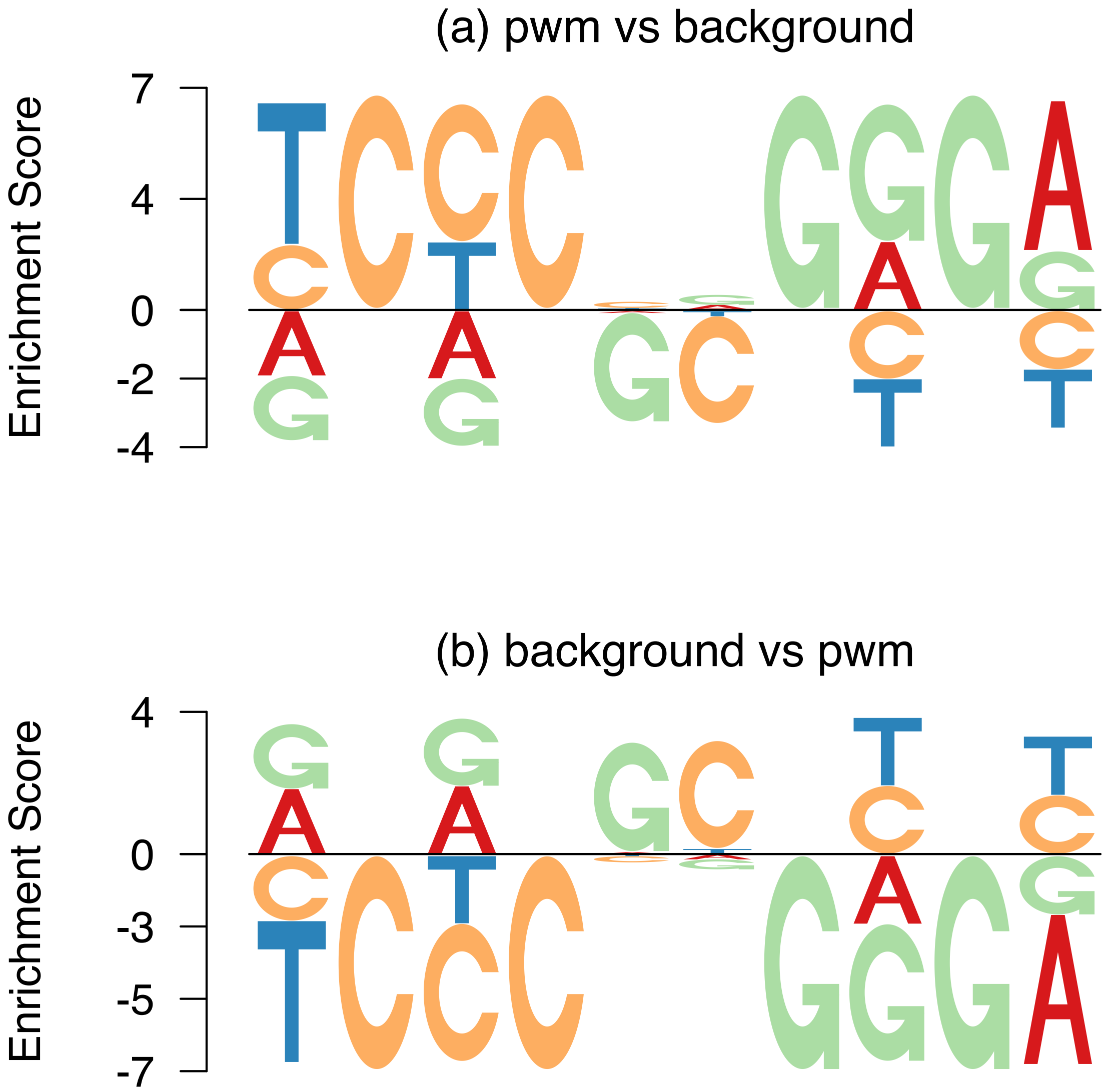


Figure S16: **Illustration of "mirror property" of *EDLogo*.** *Panel (a)*: *EDLogo* plot of the position weight matrix (PWM) of the primary discovered motif *disc1* from [63] of the EBF1 transcription factor against uniform background. *Panel (b)*:*EDLogo* plot of a uniform PWM against the PWM of EBF1 as background. That is, panels (a) and (b) are comparing the same two PWMs, but differ in which one they treat as the "background". The *EDLogo* plot obeys the mirror property, in that (b) is a mirror image of (a) (modulo the orientation of the symbols, which are translated and not reflected).

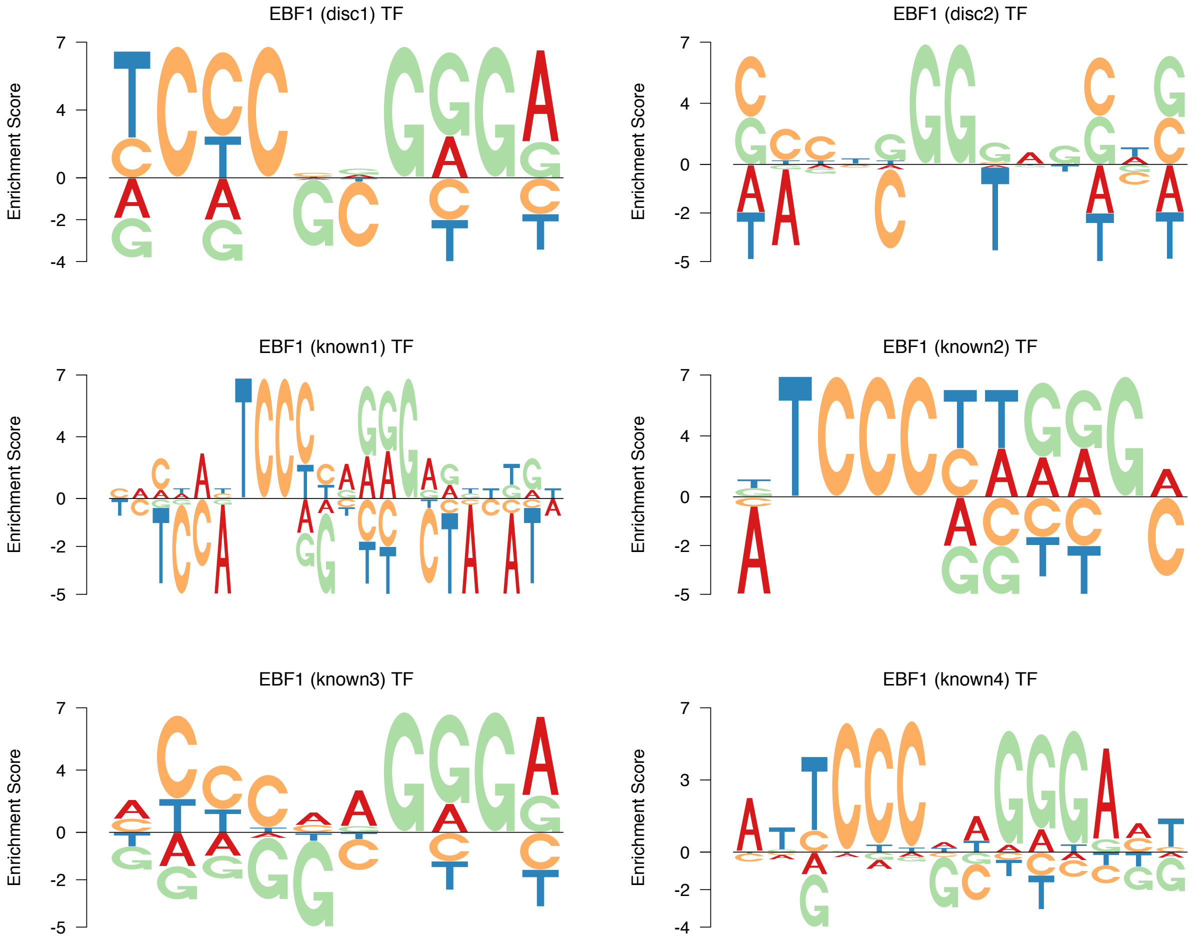


Figure S17: ***EDLogo* plots for six different motifs of the EBF1 transcription factor.** The PWMS for *known1* and *known2* come from the TRANSFAC database [135]; *known3* from the JASPAR database [103]; *known4* from [59]; *disc1* and *disc2* were discovered by the ENCODE project [63]. Three of the motifs (*known3*, *known4* and *disc1*) show depletion of G and C in the middle of the binding site.

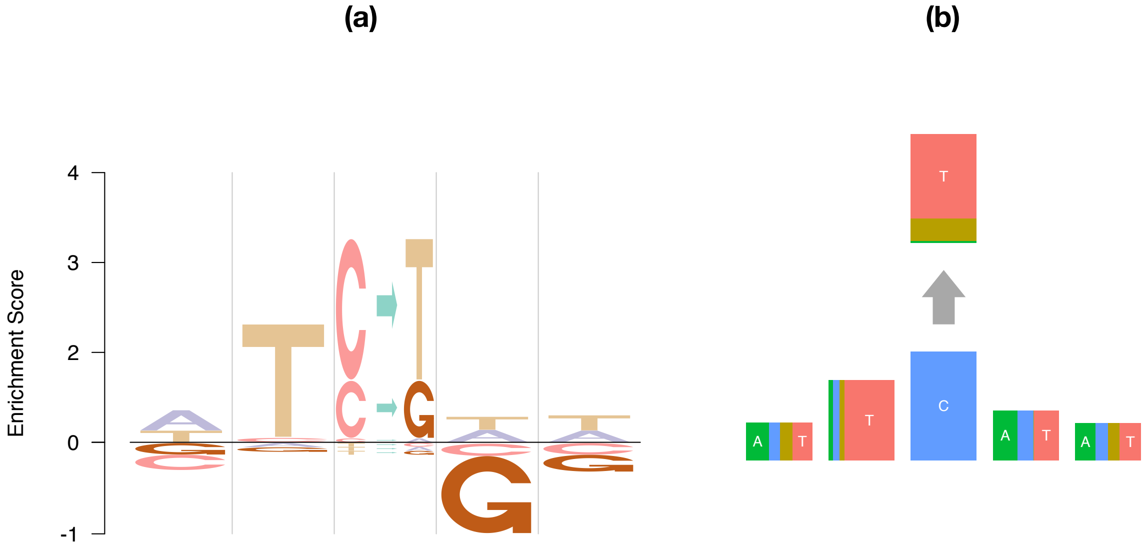


Figure S18: **Comparison of the *EDLogo* plot (a) with *pmsignature* [115] plot (b) for visualizing cancer mutational signatures.** Both plots show a signature of lymphoma B cell from [4]. The *EDLogo* plot highlights the depletion of *G* at the right flanking base more clearly than does the *pmsignature* plot. The use of strings to represent mutations in the center is arguably more intuitive than the *pmsignature* representation.

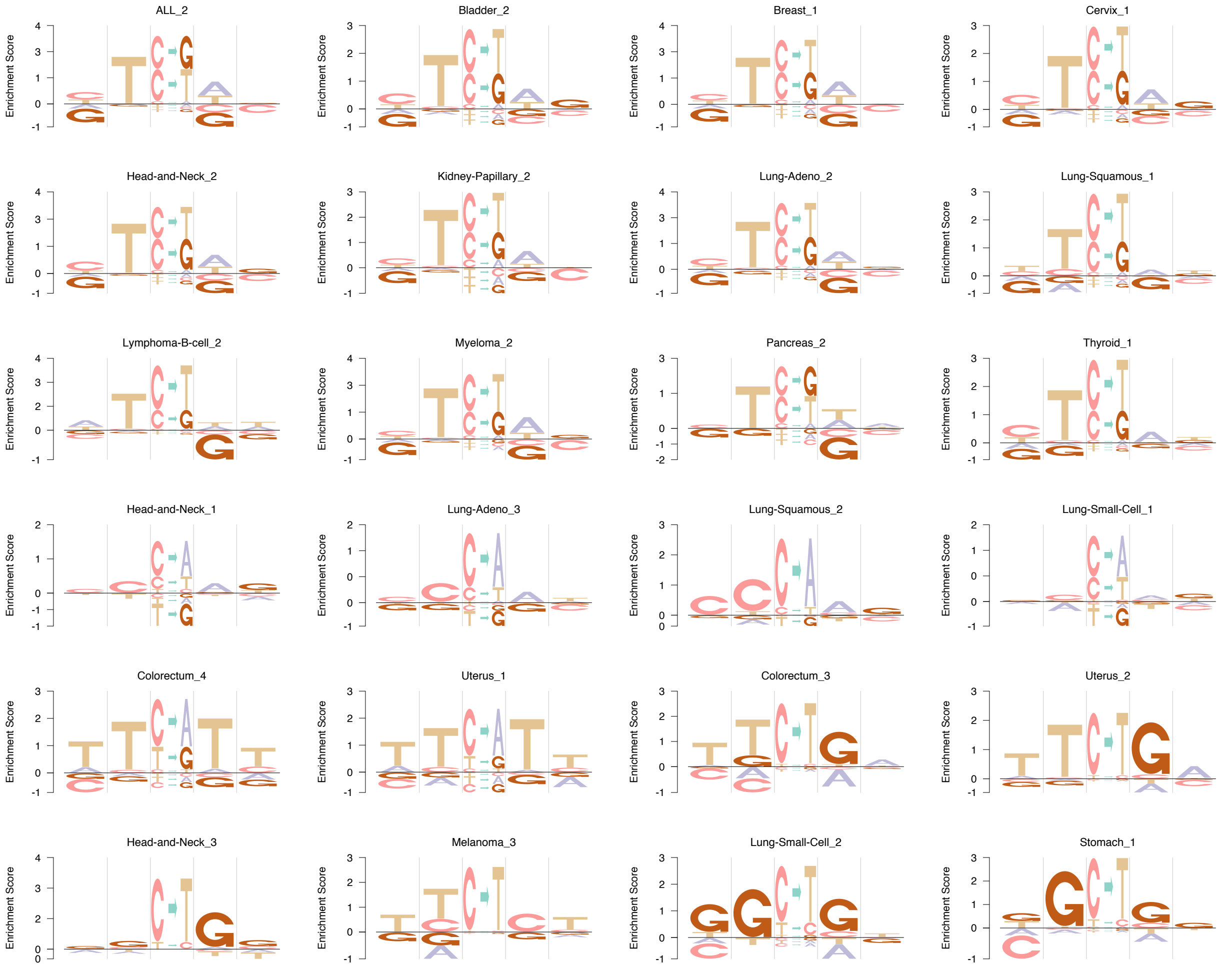


Figure S19: *EDLogo* plots for the mutation signature profiles of 24 different cancer types from [4].

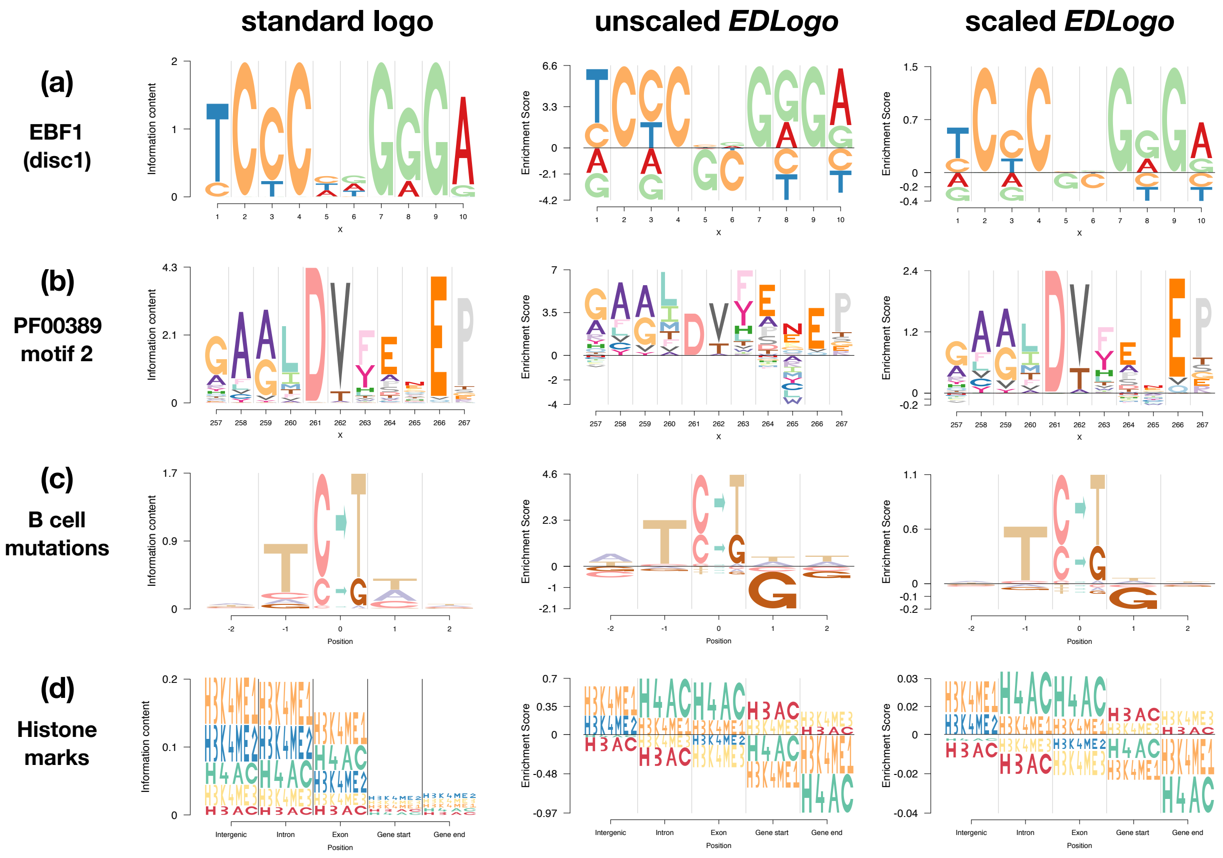


Figure S20: **Illustration of *scaled EDLogo* plot on examples from Figure 4.2.** The standard logo and unscaled *EDLogo* plots are repeated here to ease comparisons. The *scaled EDLogo* plot highlights strong enrichments more than the unscaled version and may be preferred in settings when enrichments are the primary focus.

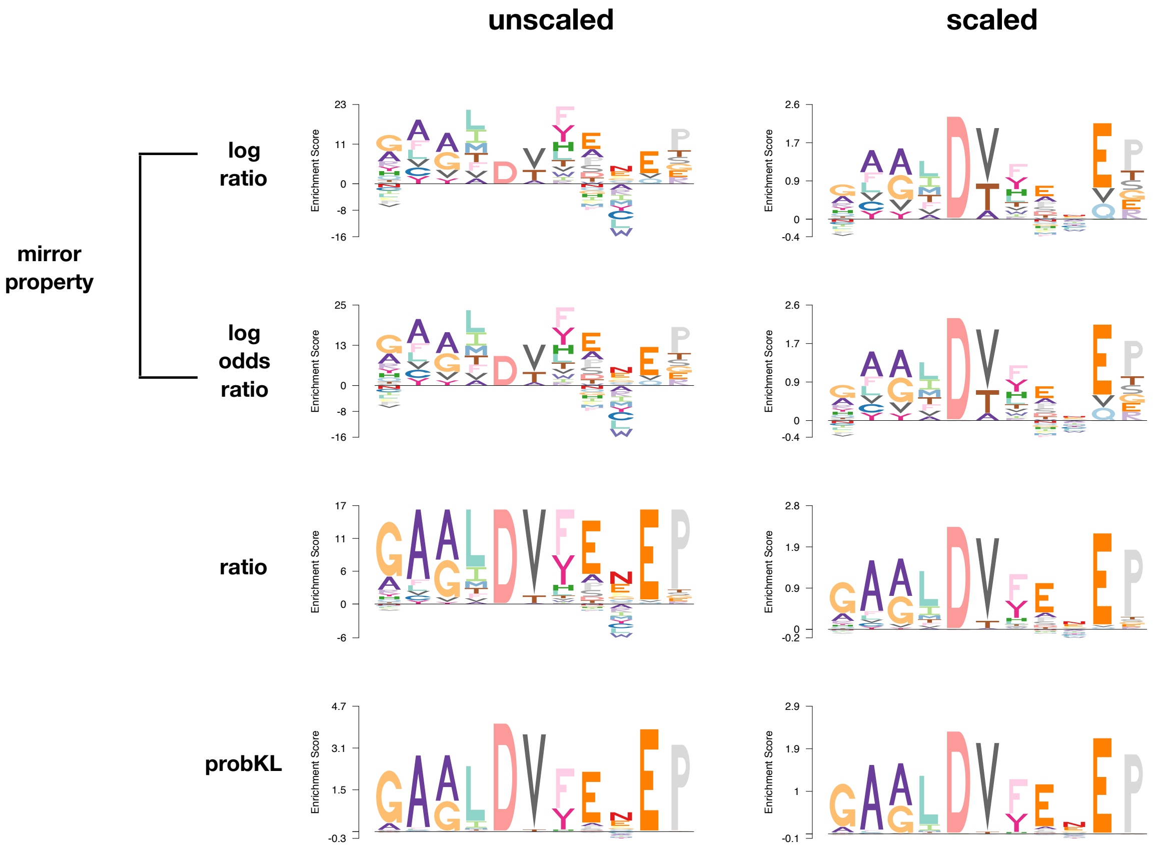


Figure S21: **Illustration of various options for *EDLogo* plot.** Each plot shows an *EDLogo* plot for a specific binding motif (Motif2 Start=257 Length=11) of the protein *D-isomer specific 2-hydroxyacid dehydrogenase, catalytic domain (IPR006139)* against a uniform background. The plots illustrate the use of several different scoring schemes (*log ratio*, *log odds ratio*, *ratio* and *probKL*) with and without scaling by the symmetric Kullback-Leibler divergence. See Supplementary Methods for details on the scoring schemes. (Note that only the *log ratio* and *log odds ratio* scoring schemes satisfy the "mirror property".)

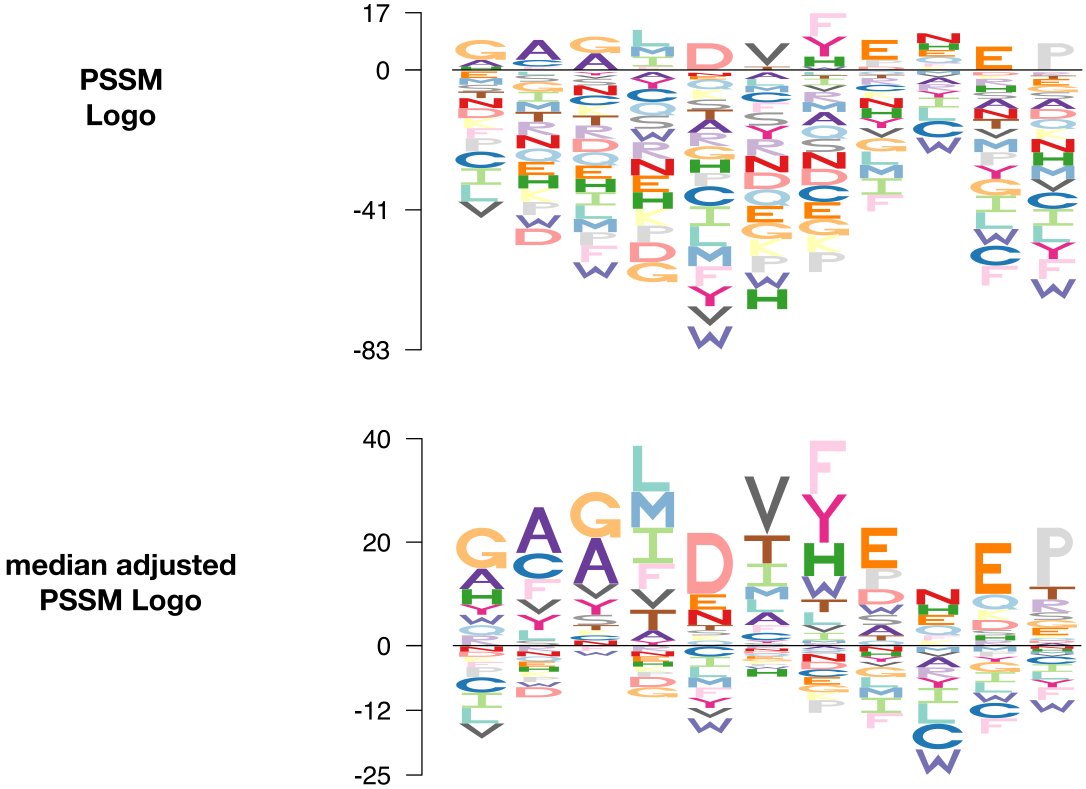


Figure S22: **Illustration of median adjustment of a position specific scoring matrix (PSSM).** The PSSM shown here is for the binding motif of the protein *D-isomer specific 2-hydroxyacid dehydrogenase, catalytic domain (IPR006139)* (Motif2,Start=257, Length=11). The median adjusted PSSM Logo (*bottom panel*) is arguably less cluttered than the non-adjusted version (*top panel*).

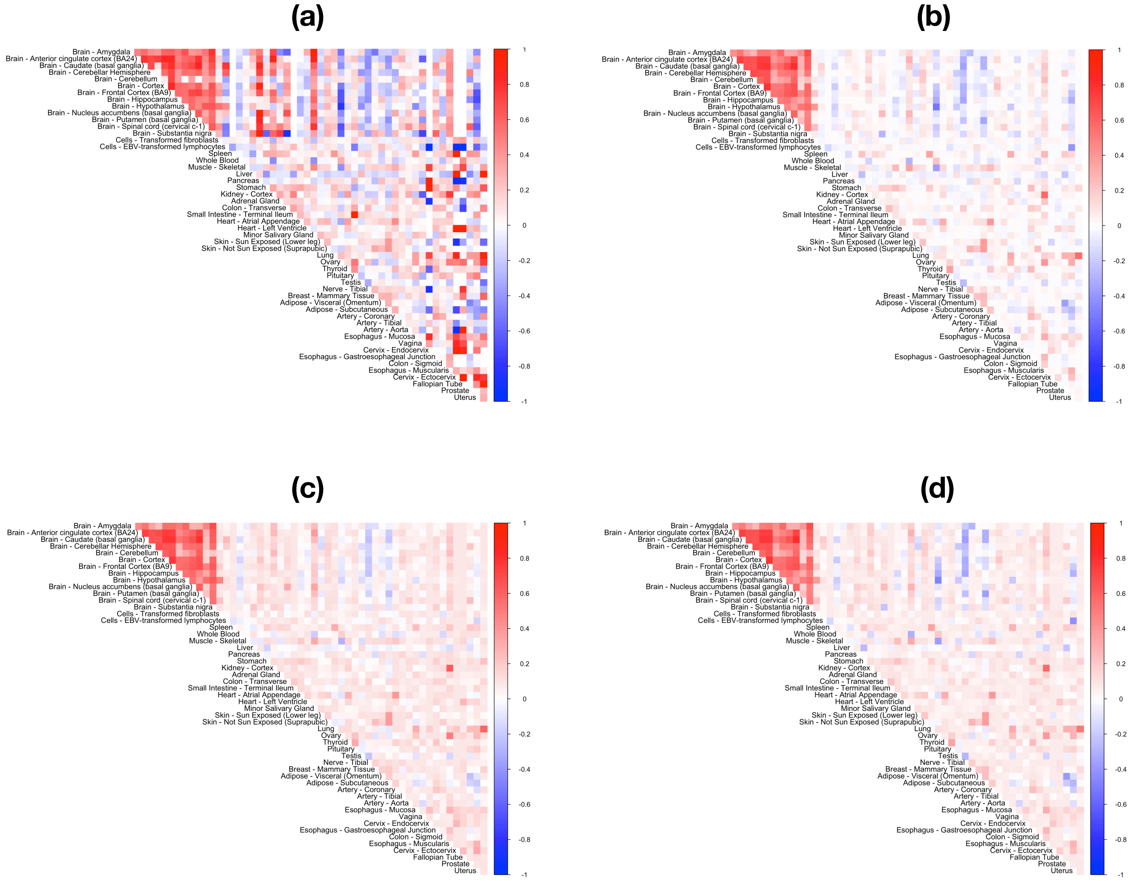


Figure S23: Application of `CorShrink` on the donor by tissue expression data for the *PLIN1* gene using different mixture model priors than the half-uniform mixture prior used in Figure 5.1. (a) presents the image plot for the pairwise sample correlation plot, (b) presents the `CorShrink` model fit with mixture normal prior centered around 0, (c) presents the `CorShrink` model fit with mixture normal prior centered around an estimated mode, (d) presents the `CorShrink` model fit using an essentially non-parametric prior (see Methods for definition of these models).

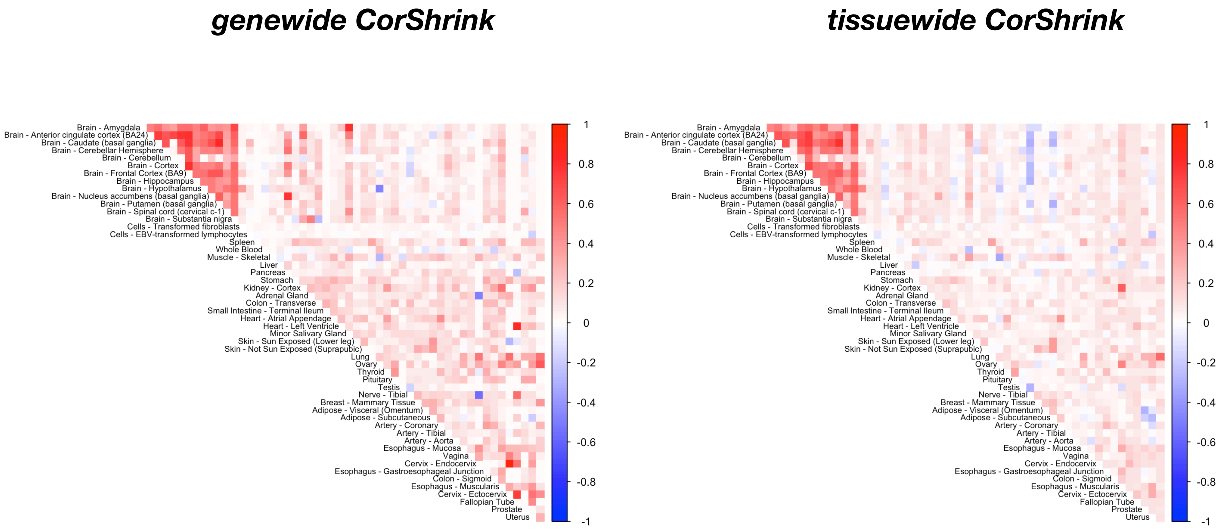


Figure S24: Image plots of the estimated correlation matrices using (a) gene-wide `CorShrink` and (b) tissue-wide `CorShrink`, both with half-uniform mixture model prior on the subject by tissue expression matrix data for the *PLIN1* gene. Both representations are visually broadly equivalent.

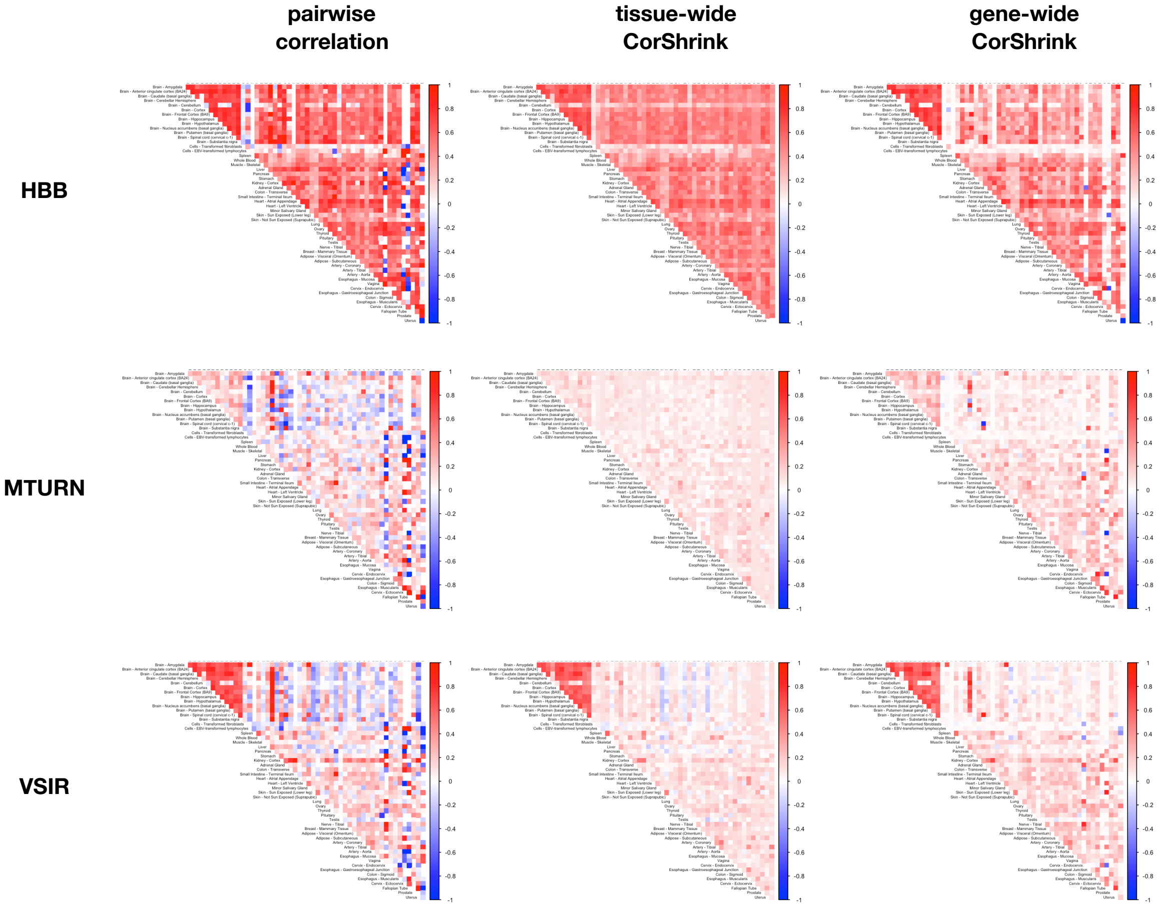
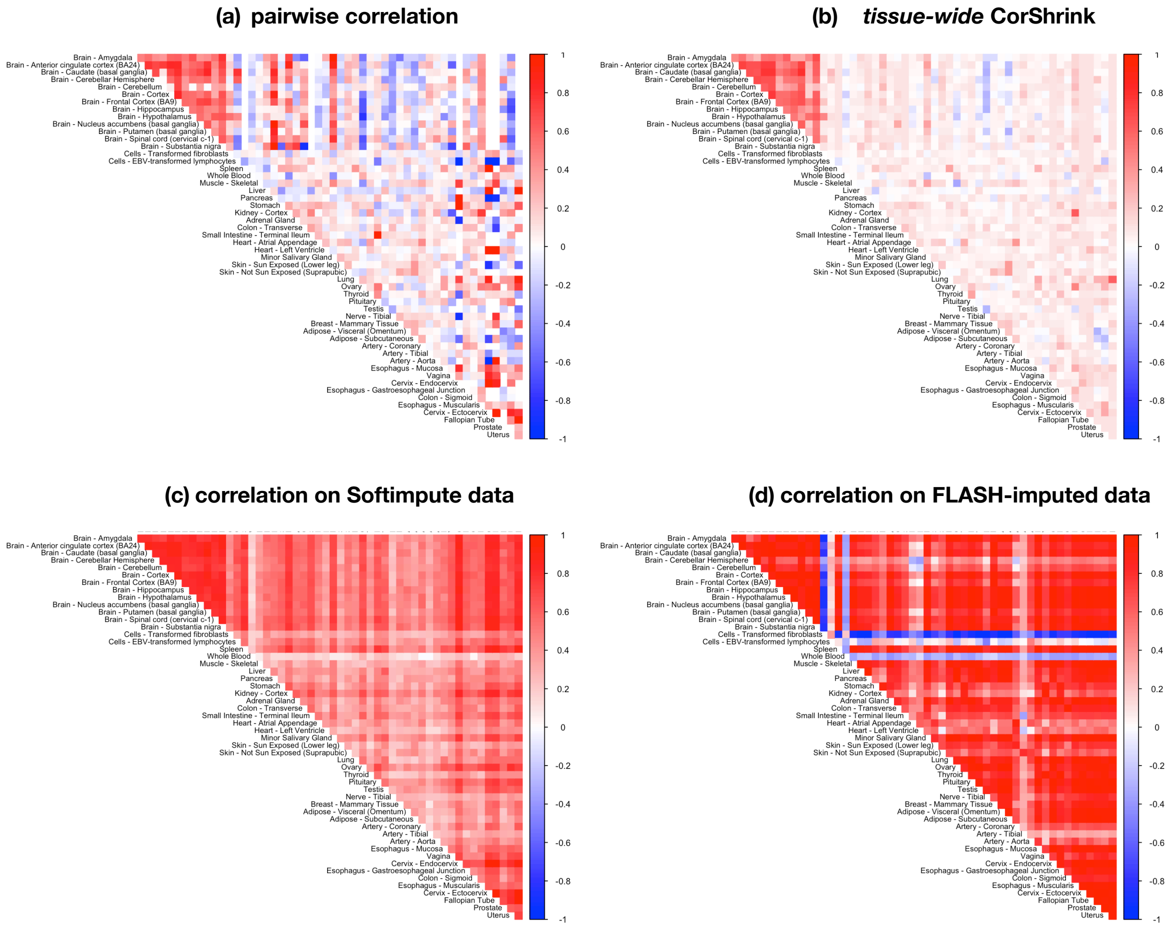


Figure S25: Image plots of the estimated correlation matrices using both tissue-wide `CorShrink` and gene-wide `CorShrink`, using half-uniform mixture model prior, on the subject by tissue expression data for three different genes - *HBB*, *MTURN* and *VSIR*. *HBB* correlation patterns are similar to *PLIN1* in Figure 5.1 with high correlation among the Brain tissues and negligible correlation among other tissues. *MTURN* exhibits low correlation across all tissues and *VSIR* exhibits high correlation across almost all the tissues.



**(a) pairwise correlation**

**(b) *tissue-wide* CorShrink**

**(c) correlation on Softimpute data**

**(d) correlation on FLASH-imputed data**

Figure S26: Comparison of the tissue-tissue correlation matrix of the log CPM expression data for *PLIN1* gene (a) and the estimated correlation matrix from our proposed `CorShrink` method (b) with respect to correlation matrices obtained after imputing the missing observations in the data matrix by SoftImpute (c) and FLASH (d). `CorShrink` seems to capture the subtle structure in the tissue tissue correlations and generate a visually more interpretable representation than the imputation mechanisms.

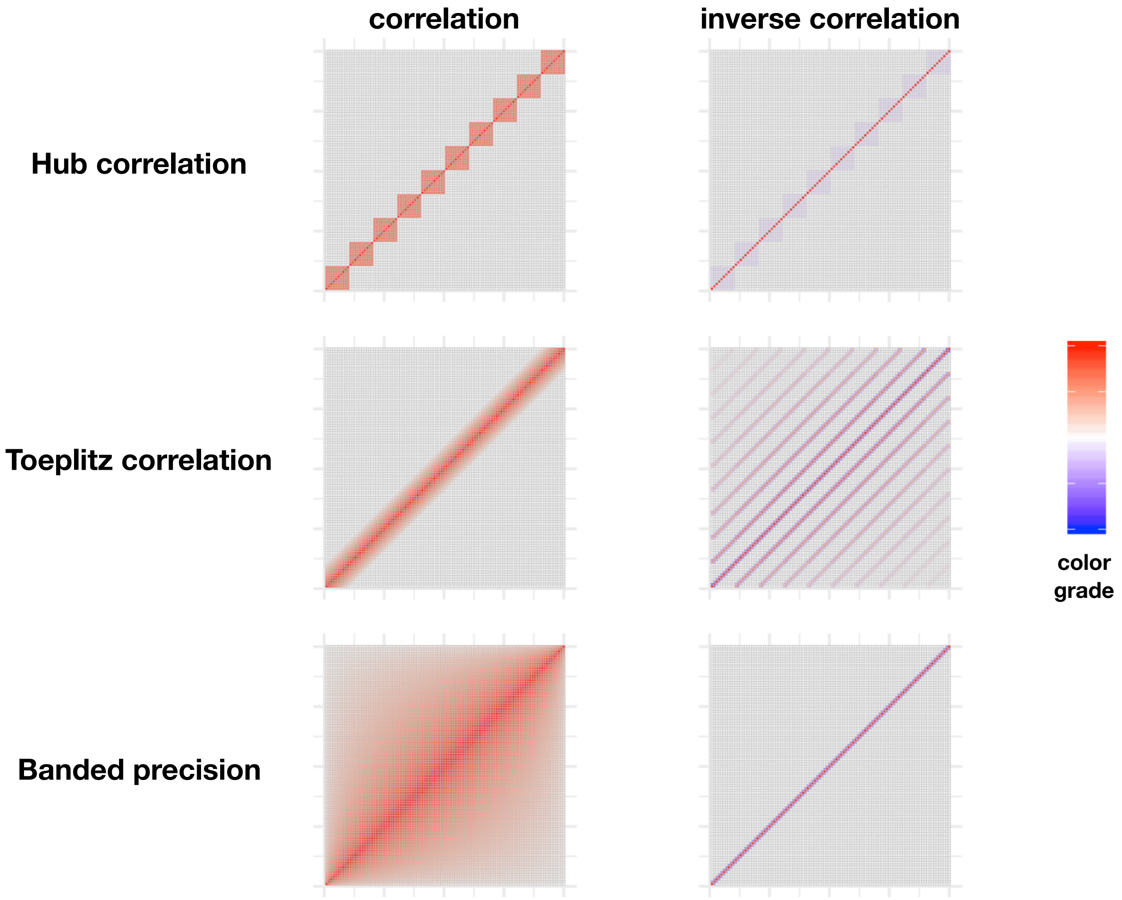


Figure S27: A demonstration of the population correlation and inverse correlation structure from which simulation studies were carried out.

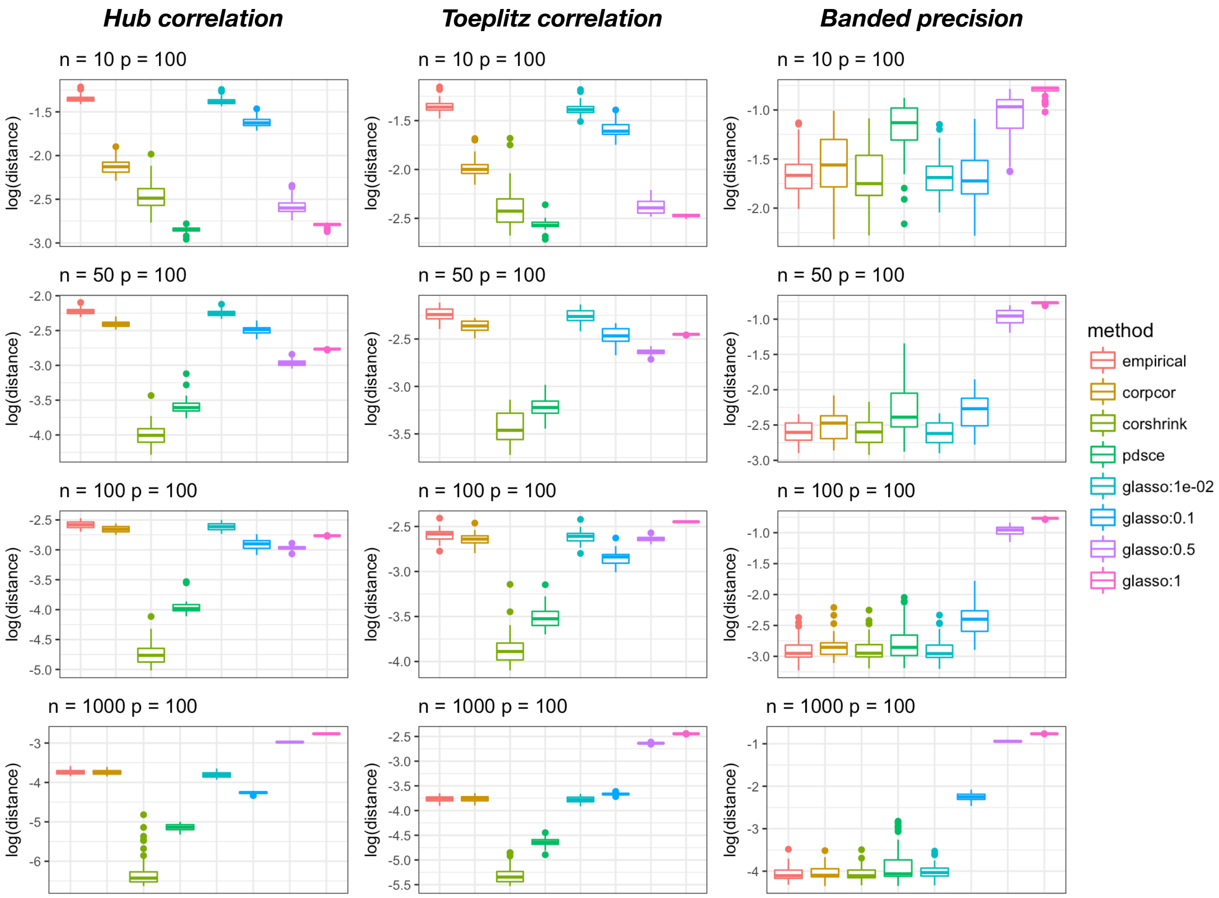


Figure S28: Box plot of the Frobenius distance between population correlation matrix and the estimated matrix from different methods - *corpcor* [106, 107],`CorShrink`, *PDSCE* [101], GLASSO [40] at different tuning parameters and the empirical pairwise correlation matrix, for different structural assumptions on the underlying population correlation - Hub structure, Toeplitz structure and a banded precision matrix, see Supplementary Figure S27. `CorShrink` outperforms the other methods for the structured/sparse covariance models (Hub and Toeplitz), with *PDSCE* being closest to `CorShrink` in performance..

# References

[1] a. 1000 Genomes Project Consortium, *An integrated map of genetic variation from 1,092 human genomes*, Nature, 491 (2012), p. 56.

[2] J. Ahn, Y. Yuan, G. Parmigiani, M. Suraokar, L. Diao, I. Wistuba, and W. Wang, *Demix: deconvolution for mixed cancer transcriptomes using raw measured data*, Bioinformatics, 29(15) (2013), pp. 1865–71.

[3] D. H. Alexander, J. Novembre, and K. Lange, *Fast model-based estimation of ancestry in unrelated individuals*, Genome research, 19 (2009), pp. 1655–1664.

[4] L. Alexandrov, G. Nik-Zainal, D. Wedge, P. Campbell, and M. Stratton, *Deciphering signatures of mutational processes operative in human cancer.*, Cell Reports, 3(1) (2013), pp. 246–259.

[5] A. Alizadeh, M. Eisen, R. Davis, C. Ma, I. Lossos, A. Rosenwald, and J. Boldrick, *Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling.*, Nature, 403(6769) (2000), pp. 503–11.

[6] M. E. Allentoft, M. Sikora, K.-G. Sjögren, S. Rasmussen, M. Rasmussen, J. Stenderup, P. B. Damgaard, H. Schroeder, T. Ahlström, L. Vinner, et al., *Population genomics of bronze age eurasia*, Nature, 522 (2015), pp. 167–172.

[7] H. Baba, K. Nakahira, N. Morita, F. Tanaka, H. Akita, and K. Ikenaka, *Gfap gene expression during development of astrocyte.*, Dev Neurosci., 19(1) (1997), pp. 14863–14868.

[8] O. Bembom, *seqlogo: Sequence logos for dna sequence alignments.* R package version 1.42.0.

[9] A. Bhattacharya and D. Dunson, *Sparse bayesian infinite factor models.*, Biometrika, 98(2) (2011), pp. 291–306.

[10] P. J. Bickel and E. Levina, *Covariance regularization by thresholding*, The Annals of Statistics, (2008), pp. 2577–2604.

[11] J. Bien and R. J. Tibshirani, *Sparse estimation of a covariance matrix*, Biometrika, 98 (2011), pp. 807–820.

[12] A. Bishara and J. Hittner, *Reducing bias and error in the correlation coefficient due to nonnormality*, Educational and psychological measurement, 75(5) (2015), pp. 785–804.

[13] A. Bishara and J. Hittner, *Confidence intervals for correlations when data are not normal*, Behavior research methods, 49(1) (2017), pp. 294–309.

[14] D. Blei and J. Lafferty, *Topic models*, In A. Srivastava and M. Sahami, editors, Text Mining: Classification, Clustering, and Applications . Chapman and Hall/CRC Data Mining and Knowledge Discovery Series, (2009).

[15] D. Blei, A. Ng, and M. Jordan, *Latent dirichlet allocation*, J. Mach. Learn. Res., 3 (2003), pp. 993–1022.

[16] A. W. Briggs, U. Stenzel, P. L. Johnson, R. E. Green, J. Kelso, K. Prüfer, M. Meyer, J. Krause, M. T. Ronan, M. Lachmann, et al., *Patterns of damage in genomic dna sequences from a neandertal*, Proceedings of the National Academy of Sciences, 104 (2007), pp. 14616–14621.

[17] H. M. Cann, C. De Toma, L. Cazes, M.-F. Legrand, V. Morel, L. Piouffre, J. Bodmer, W. F. Bodmer, B. Bonne-Tamir, A. Cambon-Thomsen, et al., *A human genome diversity cell line panel*, Science, 296 (2002), pp. 261–262.

[18] N. Coalert, K. Helsens, L. Martens, J. Vandekerckhove, and K. Gevaert, *Improved visualization of protein consensus sequences by icelogo*, Nature Methods, 6 (2009), pp. 786–787.

[19] T. G. Consortium, *The genotype-tissue expression (gtex) project.*, Nature genetics, 45(6) (2013), pp. 580–585.

[20] G. Crooks, *Weblogo: A sequence logo generator.*, Genome Research, 14 (6) (2004), pp. 1188–1190.

[21] F. Danielsson, T. James, D. Gomez-Cabrero, and M. Huss, *Assessing the consistency of public human tissue rna-seq data sets.*, Briefings in Bioinformatics, (2015).

[22] Q. Deng, D. Ramskold, B. Reinius, and R. Sandberg, *Single-cell rna-seq reveals dynamic, random monoallelic gene expression in mammalian cells.*, Science, 343(6167) (2014), pp. 193–196.

[23] K. Dey, C. Hsiao, and S. M, *Visualizing the structure of rna-seq expression data using grade of membership models.*, PLOS Genetics, 13(3) (2017), p. e1006599.

[24] K. Dey, C. Hsiao, and M. Stephens, *Visualizing the structure of rna-seq expression data using grade of membership models*, PLoS genetics, 13 (2017), p. e1006599.

[25] K. Dey, J. Hsiao, and M. Stephens, `CountClust` *: Clustering and visualizing rna-seq expression data using grade of membership models*, R package version 0.99.3, (2016).

[26] K. Dey, D. Xie, and M. Stephens, *A new sequence logo plot to highlight enrichment and depletion*, bioRxiv, p.226597, (2017).

[27] P. D'haeseleer, *How does gene expression clustering work?*, Nat Biotechnol, 23(12) (2005), pp. 1499–501.

[28] P. Diaconis and B. Efron, *Computer-intensive methods in statistics*, Scientific American, 248(5) (1983), pp. 116–131.

[29] P. Diaconis, S. Goel, and S. Holmes, *Horseshoes in multidimensional scaling and local kernel methods.*, Ann. Appl. Stat, 2(3) (2008), pp. 777–807.

[30] B. K. Duncan and J. H. Miller, *Mutagenic deamination of cytosine residues in dna*, Nature, 287 (1980), p. 560.

[31] B. Efron, *Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods*, Biometrika, 68(3) (1981), pp. 589–599.

[32] M. Eisen, P. Spellman, P. Brown, and D. Botstein, *Cluster analysis and display of genome-wide expression patterns.*, PNAS, 95(25) (1998).

[33] B. Engelhardt and M. Stephens, *Analysis of population structure: A unifying framework and novel methods based on sparse factor analysis.*, PLOS Genetics, (2010).

[34] E. Erosheva, *Latent class representation of the grade of membership model*, Seattle: University of Washington., (2006).

[35] A. Evsikov and C. De Evsikova, *Gene expression during the oocyte-to-embryo transition in mammals.*, Molecular Reproduction and Development, 76 (2009), pp. 805–818.

[36] G. Falco, S. Lee, I. Stanghellini, U. Bassey, T. Hamatani, and M. Ko, *Zscan4: a novel gene expressed exclusively in late 2-cell embryos and embryonic stem cells.*, Developmental biology, 307(2) (2007), pp. 539–550.

[37] D. Falush, M. Stephens, and J. Pritchard, *Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies.*, Genetics, 164(4) (2003), pp. 1567–87.

[38] R. Fisher, *Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population*, Biometrika, 10(4) (1915), pp. 507–521.

[39] R. Fisher, *On the probable error of a coefficient of correlation deduced from a small sample*, Metron, 1 (1921), pp. 3–32.

[40] J. Friedman, T. Hastie, and R. Tibshirani, *Sparse inverse covariance estimation with the graphical lasso*, Biostatistics, 9(3) (2008), pp. 432–441.

[41] Q. Fu, H. Li, P. Moorjani, F. Jay, S. M. Slepchenko, A. A. Bondarev, P. L. Johnson, A. Aximu-Petri, K. Prüfer, C. de Filippo, et al., *Genome sequence of a 45,000-year-old modern human from western siberia*, Nature, 514 (2014), pp. 445–449.

[42] C. Gamba, E. R. Jones, M. D. Teasdale, R. L. McLaughlin, G. Gonzalez-Fortes, V. Mattiangeli, L. Domboróczki, I. Kővári, I. Pap, A. Anders, et al., *Genome flux and stasis in a five millennium transect of european prehistory*, Nature communications, 5 (2014), p. 5257.

[43] R. Gentleman, D. Bates, B. Bolstad, and et al, *Bioconductor: a software development project.*, Technical Report, Department of Biostatistics, Harvard School of Public Health, Boston, (2003).

[44] Y. Gilad and O. Mizrahi-Man, *A reanalysis of mouse encode comparative gene expression data.*, F1000Research, 4:121 (2015).

[45] A. Ginolhac, M. Rasmussen, M. T. P. Gilbert, E. Willerslev, and L. Orlando, *mapdamage: testing for damage patterns in ancient dna sequences*, Bioinformatics, 27 (2011), pp. 2153–2155.

[46] N. Goldman and Z. Yang, *A codon-based model of nucleotide substitution for protein-coding dna sequences.*, Molecular biology and evolution, 11 (1994), pp. 725–736.

[47] T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M. Caligiuri, C. Bloomfield, and E. Lander, *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.*, Science, 286(5439) (1999), pp. 531–7.

[48] G. Guo, M. Huss, G. Tong, C. Wang, L. Sun, N. Clarke, and R. P, *Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst*, Developmental Cell, 18(4) (2010), pp. 675–685.

[49] T. Hastie and R. Mazumder, *softimpute: Matrix completion via iterative soft-thresholded svd*, R package version, 1., (2015).

[50] M. Herdin, N. Czink, H. Ozcelik, and E. Bonek, *Correlation matrix distance, a meaningful measure for evaluation of non-stationary mimo channels*, In Vehicular Technology Conference, 2005. VTC 2005-Spring. 2005 IEEE 61st, 1 (2005), pp. 136–140.

[51] M. Hernandez, P. Andres-Barquin, S. Martinez, A. Bulfone, J. Rubenstein, and M. Israel, *Enc-1: a novel mammalian kelch-related gene specifically expressed in the nervous system encodes an actin-binding protein.*, J Neurosci., 17(9) (1997), pp. 3038–51.

[52] S. Hicks, M. Teng, and R. Irizarry, *On the widespread and critical impact of systematic bias and batch effects in single-cell rna-seq data.*, BiorXiv, (2015).

[53] N. Higham, *Computing the nearest correlation matrixa problem from finance*, IMA journal of Numerical Analysis, 22(3) (2002), pp. 329–343.

[54] M. Hoffman, D. Blei, and F. Bach, *Online learning for latent dirichlet allocation*, Neural Information Processing Systems, (2010).

[55] J. Hou, A. Charters, S. Lee, Y. Zhao, M. Wu, S. Jones, M. Marra, and P. Hoodless, *A systematic screen for genes expressed in definitive endoderm by serial analysis of gene expression (sage)*, BMC Developmental Biology, 7(92) (2007), pp. 1–13.

[56] J. Hu, L. Shi, Y. Chen, X. Xie, N. Zhang, A. Zhu, J. Zheng, Y. Feng, C. Zhang, J. Xi, and H. Lu, *Differential effects of myelin basic protein-activated th1 and th2 cells on the local immune microenvironment of injured spinal cord*, Experimental Neurology, 277 (2016), pp. 190–201.

[57] D. Jaitin, E. Kenigsberg, and et al, *Massively parallel single-cell rna-seq for marker-free decomposition of tissues into cell types.*, Science, 343(6172.

[58] D. Jiang, C. Tang, and A. Zhang, *Cluster analysis for gene expression data: A survey*, Microsoft Research, (2004).

[59] A. Jolma, J. Yan, T. Whitington, J. Toivonen, K. Nitta, P. Rastas, E. Morgunova, M. Enge, M. Taipale, G. Wei, and et al., *Dna-binding specificities of human transcription factors.*, Cell, 152 (2013), pp. 327–339.

[60] H. Jónsson, A. Ginolhac, M. Schubert, P. L. Johnson, and L. Orlando, *mapdamage2. 0: fast approximate bayesian estimates of ancient dna damage parameters*, Bioinformatics, 29 (2013), pp. 1682–1684.

[61] A. P. Joseph, P. Shingate, A. K. Upadhyay, and R. Sowdhamini, *3pfdb+: improved search protocol and update for the identification of representatives of protein sequence domain families.*, Database (Oxford), bau026 (2014).

[62] A. Kamburov and a. et, *The consensuspathdb interaction database: 2013 update.*, Nucleic Acids Res, (2013).

[63] P. Kheradpour and M. Kellis, *Systematic discovery and characterization of regulatory motifs in encode tf binding experiments*, Nucleic Acids Research, (2013), pp. 1–12.

[64] C. Koch and et al., *The landscape of histone modifications across 1of the human genome in five human cell lines.*, Genome Research, 17(6) (2007), pp. 691–707.

[65] T. S. Korneliussen, A. Albrechtsen, and R. Nielsen, *Angsd: analysis of next generation sequencing data*, BMC bioinformatics, 15 (2014), p. 356.

[66] T. Lancewicki and M. Aladjem, *Multi-target shrinkage estimation for covariance matrices*, IEEE Transactions on Signal Processing, 62 (2014), pp. 6380–6390.

[67] K. Lange, *A quasi-newton acceleration of the em algorithm*, Statistica sinica, (1995), pp. 1–18.

[68] I. Lazaridis, N. Patterson, A. Mittnik, G. Renaud, S. Mallick, K. Kirsanow, P. H. Sudmant, J. G. Schraiber, S. Castellano, M. Lipson, et al., *Ancient human genomes suggest three ancestral populations for present-day europeans*, Nature, 513 (2014), p. 409.

[69] O. Ledoit and M. Wolf, *Improved estimation of the covariance matrix of stock returns with an application to portfolio selection*, Journal of empirical finance, 10 (2003), pp. 603–621.

[70] O. Ledoit and M. Wolf, *A well-conditioned estimator for large-dimensional covariance matrices*, Journal of multivariate analysis, 88 (2004), pp. 365–411.

[71] J. Leek, R. Scharpf, H. Bravo, D. Simcha, B. Langmead, W. Johnson, D. Geman, K. Baggerly, and R. Irizarry, *Tackling the widespread and critical impact of batch effects in high-throughput data.*, Nature Reviews Genetics, 11 (2010), pp. 733–39.

[72] J. Leek and J. Storey, *Capturing heterogeneity in gene expression studies by surrogate variable analysis.*, PLoS Genet, 3(9) (2007).

[73] J. Lindo, E. Huerta-Sánchez, S. Nakagome, M. Rasmussen, B. Petzelt, J. Mitchell, J. S. Cybulski, E. Willerslev, M. DeGiorgio, and R. S. Malhi, *A time transect of exomes from a native american population before and after european contact*, Nature communications, 7 (2016), p. 13175.

[74] J. Lindsay, I. Mandoiu, and C. Nelson, *Gene expression deconvolution using single-cells*, BMC bioinformatics, (2013).

[75] J. Lonsdale, J. Thomas, M. Salvatore, R. Phillips, E. Lo, S. Shad, R. Hasz, G. Walters, F. Garcia, N. Young, and B. Foster, *The genotype-tissue expression (gtex) project*, Nature genetics, 45(6) (2013), p. 580.

[76] H. Lopes and M. West, *Bayesian model assessment in factor analysis.*, Statistica Sinica, 14 (2004), pp. 41–67.

[77] H. Malmström, E. M. Svensson, M. T. P. Gilbert, E. Willerslev, A. Götherström, and G. Holmlund, *More on contamination: the use of asymmetric molecular behavior to identify authentic ancient human dna*, Molecular biology and evolution, 24 (2007), pp. 998–1004.

[78] I. Mathieson et al., *The genomic history of southeastern europe*, bioRxiv, (2017).

[79] I. Mathieson, I. Lazaridis, N. Rohland, S. Mallick, N. Patterson, S. A. Roodenberg, E. Harney, K. Stewardson, D. Fernandes, M. Novak, et al.,

*Genome-wide patterns of selection in 230 ancient eurasians*, Nature, 528 (2015), pp. 499–503.

[80] R. Mazumder, T. Hastie, and R. Tibshirani, *Spectral regularization algorithms for learning large incomplete matrices*, Journal of machine learning research, 11 (2010), pp. 2287–2322.

[81] N. Meinshausen and P. Buhlmann, *High-dimensional graphs and variable selection with the lasso*, The annals of statistics, (2006), pp. 1436–1462.

[82] T. Mikolov, K. Chen, G. Corrado, and J. Dean, *Efficient estimation of word representations in vector space*, arXiv preprint arXiv:1301.3781.

[83] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, *Distributed representations of words and phrases and their compositionality*, In Advances in neural information processing systems, (2013), pp. 3111–3119.

[84] J. Novembre and M. Stephens, *Interpreting principal component analyses of spatial population genetic variation.*, Nat Genet., 40(5) (2008), pp. 646–649.

[85] I. Olalde et al., *The beaker phenomenon and the genomic transformation of northwest europe*, bioRxiv, (2017).

[86] A. Oshlack, M. Robinsom, and M. Young, *From rna-seq reads to differential expression results.*, Genome Biology, 11:220 (2010).

[87] J. Pennington, R. Socher, and C. Manning, *Glove: Global vectors for word representation.*

[88] K. Pentchev and a. et, *Evidence mining and novelty assessment of protein-protein interactions with the consensuspathdb plugin for cytoscape.*, Bioinformatics, (2010).

[89] J. K. Pritchard, M. Stephens, and P. Donnelly, *Inference of population structure using multilocus genotype data*, Genetics, 155 (2000), pp. 945–959.

[90] K. Prüfer, F. Racimo, N. Patterson, F. Jay, S. Sankararaman, S. Sawyer, A. Heinze, G. Renaud, P. H. Sudmant, C. De Filippo, et al., *The complete genome sequence of a neandertal from the altai mountains*, Nature, 505 (2014), p. 43.

[91] F. Putkey, T. Cramer, M. Morphew, A. Silk, R. Johnson, and J. Mclntosh, *Unstable kinetochore-microtubule capture and chromosomal instability following deletion of cenp-e.*, Developmental cells, 3(3) (2002), pp. 351–365.

[92] W. Qiao, G. Quon, E. Csaszar, M. Yu, Q. Morris, and P. Zandstra, *Pert: A method for expression deconvolution of human blood samples from varied microenvironmental and developmental conditions*, PLoS Comput Biol, 8(12) (2012).

[93] G. Quon, S. Haider, A. Deshwar, A. Cui, P. Boutros, and Q. Morris, *Computational purification of individual tumor gene expression profiles leads to significant improvements in prognostic prediction*, Genome Med., 5(3) (2013), p. 29.

[94] M. Rasmussen, X. Guo, Y. Wang, K. E. Lohmueller, S. Rasmussen, A. Albrechtsen, L. Skotte, S. Lindgreen, M. Metspalu, T. Jombart, et al., *An aboriginal australian genome reveals separate human dispersals into asia*, Science, 334 (2011), pp. 94–98.

[95] G. Renaud, V. Slon, A. T. Duggan, and J. Kelso, *Schmutzi: estimation of contamination and endogenous mitochondrial consensus calling for ancient dna*, Genome biology, 16 (2015), p. 224.

[96] D. Repsilber, S. Kern, A. Telaar, G. Walzl, G. Black, J. Selbig, S. Parida, S. Kaufmann, and M. Jacobsen, *Biomarker discovery in heterogeneous tissue samples -taking the in-silico deconfounding approach*, BMC bioinformatics, 11(1) (2010), pp. 27+.

[97] N. Rohland, E. Harney, S. Mallick, S. Nordenfelt, and D. Reich, *Partial uracil–dna–glycosylase treatment for screening of ancient dna*, Phil. Trans. R. Soc. B, 370 (2015), p. 20130624.

[98] N. Rosenberg, *Algorithms for selecting informative marker panels for population assignment.*, J Comput Biol, 12(9) (2005), pp. 1183–201.

[99] N. A. Rosenberg, J. K. Pritchard, J. L. Weber, H. M. Cann, K. K. Kidd, L. A. Zhivotovsky, and M. W. Feldman, *Genetic structure of human populations*, science, 298 (2002), pp. 2381–2385.

[100] J. Rossant and P. Tam, *Blastocyst lineage formation, early embryonic assymmetries and axis patterning in the mouse.*, Development, 136(5) (2009), pp. 701–13.

[101] A. J. Rothman, *Positive definite estimators of large covariance matrices*, Biometrika, 99 (2012), pp. 733–740.

[102] A. J. Rothman, E. Levina, and J. Zhu, *Generalized thresholding of large covariance matrices*, Journal of the American Statistical Association, 104 (2009), pp. 177–186.

[103] A. Sandelin, A. Wynand, P. Engstrom, W. W.W., and B. Lenhard, *Jaspar: an open-access database for eukaryotic transcription factor binding profiles*, Nucleic Acids Research, 32 (Database issue) (2004), pp. D91–D94.

[104] S. Sawyer, J. Krause, K. Guschanski, V. Savolainen, and S. Pääbo, *Temporal patterns of nucleotide misincorporations and dna fragmentation in ancient dna*, PloS one, 7 (2012), p. e34131.

[105] E. Scarano, M. Iaccarino, P. Grippo, and E. Parisi, *The heterogeneity of thymine methyl group origin in dna pyrimidine isostichs of developing sea urchin embryos.*, Proc. Natl. Acad. Sci., 57 (5) (1967), p. 1394400.

[106] J. Schäfer and K. Strimmer, *An empirical bayes approach to inferring large-scale gene association networks*, Bioinformatics, 21 (2004), pp. 754–764.

[107] J. Schäfer and K. Strimmer, *A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics*, Statistical applications in genetics and molecular biology, 4 (2005).

[108] C. Schaniel, Y. Ang, K. Ratnakumar, C. Cormier, T. James, E. Bernstein, I. Lemischka, and P. Paddison, *Smarcc1/baf155 couples self-renewal gene repression with changes in chromatic structure in mouse embroynic stem cells.*, Stem cells, 27(12) (2009), pp. 2979–91.

[109] T. D. Schneider and R. Stephens, *Sequence logos: a new way to display consensus sequences*, Nucleic Acids Research, 18 (20) (1990), pp. 6097–6100.

[110] R. Schwartz and S. Shackney, *Applying unmixing to gene expression data for tumor phylogeny inference*, BMC bioinformatics, 11(1) (2010), pp. 42+.

[111] K. Shameer, P. Nagarajan, K. Gaurav, and R. Sowdhamini, *3pfdb - a database of best representative pssm profiles (brps) of protein families generated using a novel data mining approach.*, BioData Min., 2(1) (2009), p. 8.

[112] Á. Shapiro and M. Hofreiter, *A paleogenomic perspective on evolution and gene function: new insights from ancient dna*, Science, 343 (2014), p. 1236573.

[113] J.-C. Shen, W. M. Rideout III, and P. A. Jones, *The rate of hydrolytic deamination of 5-methylcytosine in double-stranded dna*, Nucleic acids research, 22 (1994), pp. 972–976.

[114] S. Shen-Orr, R. Tibshirani, P. Khatri, D. Bodian, F. Staedtler, N. Perry, T. Hastie, M. Sarwal, M. Davis, and A. Butte, *Cell type specific gene expression differences in complex tissues*, Nature Methods, 7(4) (2010), pp. 287–289.

[115] Y. Shiraishi, G. Tremmel, S. Miyano, and M. Stephens, *A simple model-based approach to inferring and visualizing cancer mutation signatures.*, PLoS Genetics, 11(12) (2015), p. e1005657.

[116] P. Skoglund, H. Malmström, A. Omrak, M. Raghavan, C. Valdiosera, T. Günther, P. Hall, K. Tambets, J. Parik, K.-G. Sjögren, et al., *Genomic diversity and admixture differs for stone-age scandinavian foragers and farmers*, Science, 344 (2014), pp. 747–750.

[117] P. Skoglund, B. H. Northoff, M. V. Shunkov, A. P. Derevianko, S. Pääbo, J. Krause, and M. Jakobsson, *Separating endogenous ancient dna from modern day contamination in a siberian neandertal*, Proceedings of the National Academy of Sciences, 111 (2014), pp. 2229–2234.

[118] O. Stegle, L. Parts, M. Piipari, J. Winn, and R. Durbin, *Using probabilistic estimation of expression residuals (peer) to obtain increased power and interpretability of gene expression analyses.*, Nat Protoc., 7(3) (2012), pp. 500–07.

[119] M. Stephens, *False discovery rates: a new deal*, Biostatistics, 18 (2) (2016), pp. 275–294.

[120] G. D. Stormo, *Dna binding sites: representation and discovery*, Bioinformatics, 16 (1) (2000), pp. 16–23.

[121] M. Taddy, *On estimation and selection for topic models*, in International Conference on Artificial Intelligence and Statistics, 2012, pp. 1184–1193.

[122] G. Tan and B. Lenhard, *Tfbstools: an r/bioconductor package for transcription factor binding site analysis*, Bioinformatics, 32(10) (2016), pp. 1555–1556.

[123] F. Tang, C. Barbacioru, Y. Wang, E. Nordman, C. Lee, and et al, *mrna-seq whole-transcriptome analysis of a single cell*, Nature Methods, 6 (2009), pp. 377–382.

[124] M. Thompson and C. Wu, *mygene: Access mygene.info services.*, R package version 1.2.3, (2015).

[125] M. Thomsen and M. Nielsen, *Seq2logo: a method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion*, Nucleic Acids Research, 40 (2012), pp. 281–287.

[126] A. Touloumis, *Nonparametric stein-type shrinkage covariance matrix estimators in high-dimensional settings*, Computational Statistics & Data Analysis, 83 (2015), pp. 251–261.

[127] C. Trapnell, *Defining cell types and states with single-cell genomics.*, Genome Res., 25 (2015), pp. 1491–1498.

[128] L. Van der Maaten, *Accelerating t-sne using tree-based algorithms.*, J. Mach. Learn. Res., (2014), pp. 3221–3245.

[129] L. Van der Maaten and G. Hinton, *Visualizing high-dimensional data using t-sne.*, J. Mach. Learn. Res., (2008), pp. 2579–2605.

[130] O. Wagih, *Rweblogo: plotting custom sequence logos.* R package version 1.0.3.

[131] O. Wagih, *ggseqlogo: a versatile r package for drawing sequence logos*, Bioinformatics, btx469 (2017).

[132] N. Wang, T. Gong, R. Clarke, L. Chen, I. Shih, Z. Zhang, D. Levine, J. Xuan, and Y. Wang, *Undo: a bioconductor r package for unsupervised deconvolution of mixed gene expressions in tumor samples*, Bioinformatics, 31(1) (2015), pp. 137–9.

[133] W. Wang and M. Stephens, *Empirical bayes matrix factorization*, aRxiv, (2018), p. 1802.06931.

[134] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag New York, 2009.

[135] E. Wingender, X. Chen, R. Hehl, and et al., *Transfac: an integrated system for gene expression regulation.*, Nucleic Acids Research, 28(1) (2000), pp. 316–319.

[136] D. Witten, J. Friedman, and N. Simon, *New insights and faster computations for the graphical lasso*, Journal of Computational and Graphical Statistics, 20(4) (2011), pp. 892–900.

[137] D. Witten and R. Tibshirani, *A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis.*, Biostatistics, 10(3) (2009), pp. 515–534.

[138] S. Yoon, E. Kim, Y. Kim, H. Lee, K. Kim, J. Bae, and K. Lee, *Role of bcl2-like 10 (Bcl2l10) in regulating mouse oocyte maturation*, Biology of Reproduction, 81(3) (2009), pp. 497–506.

[139] X. Zhao and et al., *Jaspar 2013: An extensively expanded and updated open-access database of transcription factor binding profiles*, TBA, TBA (2013), p. TBA.