Journal of Translational Medicine

RESEARCH

Open Access

Generative and integrative modeling for transcriptomics with formalin fixed paraffin embedded material

Eliseos J. Mucaki[1], Wenhan Zhang[1], Aryamaan Saha[2], Sabina Trebinjac[3,4], Sharon Nofech-Mozes[4,5], Eileen Rakovitch[3,4], Vanessa Dumeaux[6,7] and Michael T. Hallett[1,7*]

Abstract

Background Formalin-fixed paraffin embedded (FFPE) samples suffer from the degradation of nucleic acids, a problem that becomes particularly acute with samples stored for extended periods. It remains challenging to profile FFPE using high-throughput sequencing technologies including RNA-sequencing, and the resulting FFPE RNA-seq (fRNA-seq) data has a high rate of transcript dropout, a property shared with single cell RNA-seq. Transcript counts also have high variance and are prone to extreme values, together making downstream analyses extremely challenging.

Methods We introduce the PaRaffin Embedded Formalin-FixEd Cleaning Tool (PREFFECT), a probabilistic framework for the analysis of fRNA-seq data. PREFFECT uses generative models to fit distributions to observed expression counts while adjusting for technical and biological variables. The framework can exploit multiple expression profiles generated from matched tissues for a single sample (e.g., a tumor and morphologically normal tissue) in order to stabilize profiles and impute missing counts. PREFFECT can also leverage sample-sample adjacency networks that assist graph attention mechanisms to identify the most informative correlations in the data.

Results We evaluated the distribution of transcript counts across a compendium of fRNA-seq datasets, finding the negative binomial distribution best fits the data with little evidence supporting zero-inflated extensions. We use this knowledge in the design of PREFFECT. We show that PREFFECT can accurately impute missing values from fRNAseq count matrices and adjust for batch effects. The inclusion of sample-sample adjacency networks and multiple tissues were shown to enhance sample clustering.

Conclusions The vast majority of studies to date contain at most a few hundred profiles, making it challenging to correctly infer good statistical fits for each transcript especially in complex cohorts, given the noisy, incomplete and heterogeneous nature of the data. The integrative and generative approach of PREFFECT provides better and more specific model fits than generic bulk RNA-seq tools, especially when more advanced PREFFECT models provide matched profiles are included in the analysis. The transformed data can be directly used with many well-established tools for downstream analysis tasks, empowering its use in clinical biomarker studies and diagnostics.

*Correspondence:
Michael T. Hallett
michael.hallett@uwo.ca

Full list of author information is available at the end of the article

Keywords Formalin fixed paraffin embedded, RNA-sequencing, Generative modeling, Graph attention networks

Background

Formalin-Fixed Paraffin Embedded (FFPE) material has long been used in histopathology to store samples in a manner which preserves cellular structure and tissue morphology, and has been exploited in diagnostic pathology since at least 1991 [1]. Its ease of storage and cost-effectiveness has enabled the pathology of health and disease to be cataloged with over one billion archival samples available worldwide [2–4]. FFPE constitutes a significant resource for retrospective clinical research and prospective studies, potentially eliminating the need for the collection of fresh frozen specimens. However, since it remains challenging to profile and analyze FFPE samples with modern high-throughput technologies, this resource has been under-utilized. Our goal here is to facilitate the analysis of FFPE-based RNA-sequencing studies (fRNA-seq).

It is well-established that FFPE-harvested material can be of sufficient size and quality for use in modern -omic projects [5–9] including RNA-seq [10–17]. Nevertheless, it is also well-recognized that molecular profiles of FFPE samples generated by -omic technologies including next-generation sequencing contain significant error and bias, likely in large part due to degradation and modification introduced along each step of FFPE handling. For example, during fixation, formalin causes the formation of methylene bridges which alters the structure of nucleic acids (and proteins) [18], and results in fragmentation and mutations [19]. The subsequent dehydration process causes denaturation of nucleic acids and proteins, which may reduce RNA stability after renaturation. Heat, modulation of pH levels and endogenous enzymes (e.g., nucleases) during de-crosslinking can cause further nucleic acid damage particularly to labile RNA [20]. The rate of degradation of nucleic acids is dependent on time, the fixation process used, and storage conditions [21–23]. Extraction can cause damage in several additional ways [16, 24–26].

Several studies comparing profiles of fresh frozen samples and their matched FFPE embeddings confirm that the damaging agents together induce significant degeneration [27]. For example, Jacobsen and colleagues [28] observed that RNA extracted from FFPE tissue had a median RNA integrity number (a quality measure based on the ratio of ribosomal RNA peaks and the overall RNA degradation pattern) of 2.5 and a DV200 (the percentage of RNA fragments that are greater than 200 nucleotides in length) of 48% compared to RNA extracted from fresh frozen tissue with a median integrity number of 8.1 and DV200 of 97%, representing nearly a two-fold degradation. In general, transcript counts are generally reduced in fRNA-seq profiles compared to matched fresh frozen RNA-seq profiles [5, 9]. Missing RNA species result in what is referred to as dropout or *zero counts*. Several efforts have observed an elevation in the number of zero counts in fRNA-seq profiles [9, 29]. Fragmentation of transcripts can lead to extremely high transcript counts due to several different mechanisms [30], and mutations can reduce the rate of successful mapping of reads to genomic loci [5]. Although there have been advancements in the protocols for nucleic acid extraction and preparation [17, 31], these problems persist in even the most recent datasets.

Given the degraded nature of FFPE samples, it is important to understand the distributional properties of fRNA-seq data, and to use this information to build fRNA-seq-specific normalizations, transformations and de-noising methods. To date, such efforts lag behind other types of transcriptome profiling, including single cell RNA-seq (scRNA-seq) and bulk RNA-seq. The vast majority of fRNA-seq datasets re-purpose bulk RNA-seq analysis pipelines, although our effort here shows that this practice may not be optimal. The only fRNA-specific framework is MIXnorm from Yin and colleagues [32, 33] which normalizes transcript counts using a special mixture model.

Our effort starts by characterizing the distributional qualities of the fRNA-seq data. In particular, we provide evidence that it is well-modeled by the negative binomial (NB) distribution, a property shared with bulk and scRNA-seq data and for which a considerable number of downstream applications (e.g. differential expression) are built upon. We also identify aspects of fRNA-seq that are distinct from sc- or bulk RNA-seq. One obvious difference is that FFPE studies are much smaller than most scRNA-seq studies: the estimated 1500 fRNA-seq datasets in international repositories (e.g., ENA, SRA, dbGaP) typically contain on the order of 10^2 samples, while most scRNA-seq studies contain on the order of $10^4 - 10^6$ samples.

Using the inferred distributional information, we introduce the PaRaffin Embedded Formalin-FixEd Cleaning Tool (PREFFECT), a probabilistic model for fRNA-seq data. PREFFECT uses a series of generative models to re-express observed transcript counts using either an NB distribution or a zero-inflated extension alongside metadata corresponding to known technical and biological effects in the data. Similar approaches have been explored in the context of scRNA-seq data analysis [17, 34–51]. PREFFECT contains a series of conditional variational autoencoders (cVAEs) which allow multiple tissues to be considered simultaneously in addition to graph

attention mechanisms [47, 52] which assist by highlighting information in matched samples which may assist during model training, an important capacity when high dropout rates are observed. This transformed data can be directly used for differential expression analyses, survival analyses and other common downstream tasks. Using publicly available datasets, we show how the adjusted count data from PREFFECT improves sample clustering and classification.

Methods

Statistical analyses were performed using `Python` version 3.9 [53] and `R` version 4.4 [54]. Hierarchical clustering was performed using the `seaborn clustermap` function with average distance linkage and Euclidean distance. UMAP was performed using the `umap-learn` [55] package in SCANPY [56] ($\mathrm{min\_dist} = 0.3$). We use the mean and dispersion parameterization of the negative binomial (NB) distribution for convenience, and π is used to refer to the drop-out probability in zero-inflated versions of distributions. The *k*-BET measure of cluster mixing was computed using the Python package `scib` and the adjusted rand index (ARI) of cluster purity was computed using `scikit-learn`.

A compendium of fRNA-seq datasets

We collected a large number of fRNA-seq datasets from public repositories in order to characterize their statistical properties and then later to test the capacity of PREFFECT. Datasets were included only when raw count information was available. All datasets with GSE identifiers were obtained from the Gene Expression Omnibus `https://www.ncbi.nlm.nih.gov/geo/`. The Met astatic Breast Cancer (TMBC) dataset [57] was available through cBioPortal [58]. The Sunnybrook cohorts are not currently publicly available but are pending publication. We collected clinicopathological and technical variables (such as batch identification number) whenever possible. The datasets vary across tissue, cell type, age of the cohort and profiling technique (Table S2). Transcripts with zero counts across all samples were removed. Since the PAM50 subtype for samples was not provided in most of the breast cancer FFPE datasets, we estimated it using the PAM50 classifier [59].

Characterizing the statistical properties of fRNA-seq datasets

We require an unbiased approach to determine which statistical distribution best fits transcripts across the fRNA-seq compendium. We used only a "mild" trimmed mean approach to mitigate the influence of outliers (set to 1%), although this was not used when comparing the NB versus zero-inflated NB (ZINB) distributions, as increasing the trimmed mean percentage increases the number of zeros that are removed and therefore could result in bias against zero-inflated distributions.

Expression data was fit to six distributions using two different methods. The Akaike information criterion (AIC) measures the goodness of fit as $2k - 2ln(L)$ where k is the number of parameters of the distribution and L is the likelihood of the model given the observed data. All non-zero inflated distributions were fitted using the `StatsModels` package [60]. For zero-inflated models, the AIC and distributional parameters were estimated using the `minimize` function from `scipy` [61] to optimize the negative log-likelihood; this was repeated over a range of initial dropout values π to minimize the negative log-likelihood. The D statistic of the Kolmogorov–Smirnov (KS) test was computed between the empirical cumulative distribution function for each transcript in every dataset and the cumulative distribution function of each reference distribution. The KS test was used to ensure that differences in the number of parameters between the six distributions did not unduly affect the results. Here again a 1% trimmed mean was applied to eliminate extreme outliers.

Generation of synthetic datasets

Synthetic datasets provide a convenient means to investigate technical correctness, model limits and parameter optimization for the PREFFECT generative models. Count matrices were generated across a range of mean and dispersion values for the NB distribution. For each pair $(\mu, \theta) \in \{50, 100, 500, 1000\} \times \{0.01, 0.1, 1, 2, 5, 10, 100\}$ a matrix was formed across the $N = 1000$ transcripts and $M = 1000$ samples using variates from $NB(\mu, \theta)$.

To investigate the role of the dropout rate π in parameter estimation, we constructed a synthetic dataset where the count R_g for each transcript g in each sample is formed as follows:

$$\begin{aligned} \mu_g &\sim log\ \mathcal{N}(loc = 100, scale = 0.1), \\ R_g &\sim NB(\mu_g, \theta = 1). \end{aligned}$$

In several places, we require count data generated by a ZINB process. Toward this end, each transcript in the dataset was assigned a dropout value $\pi \sim U(0\,..\,0.8)$. Then π samples across the transcript were randomly selected and set to 0.

Creation of pseudo-synthetic samples

Although contemporary fRNA-seq datasets are often small (< 500 samples), they are of sufficient size to infer important distributional properties and parameters. For each target transcript, we use its observed frequency in the original dataset and a chosen library size to form synthetic counts following the standard way to generate NB variates (e.g. Fig. 3C). This method allows us to better

ensure that samples have similar library sizes. We can use these to generate as many *pseudo-synthetic* samples as we require to train our models. When required, patient subtype was estimated in the pseudo-synthetic count data using the PAM50 classifier [59] after 1% mean trimming and a variance stabilizing transformation to the counts. A sample-sample adjacency matrix was constructed by placing an edge between two samples of the same subtype. To evaluate the contribution of the adjacency matrix to imputation, we constructed a null adjacency matrix by randomly permuting the edges in the matrix.

Fundamentals of all PREFFECT models

There are three PREFFECT models: simple, single and full (Table S1B). All models require an unadjusted count matrix $X \in \mathbb{Z}_{>0}^{M \times N}$ for the target tissue, where M is the number of samples and N is the number of transcripts. All models can accommodate metadata (e.g., batch number, DV200, percent duplicates), and patient clinicopathological variables (e.g., grade, stage of a tumor). PREFFECT uses these variables to adjust the count data and for conditional inference [62] to adjust the raw data. Missing transcript count data is assumed to be a zero count. All other conditional variables must be fully specified.

The simple PREFFECT model

A more technical exposition of the model is given in the Supplementary Information and depicted in Fig. S3A, B. The encoder produces an estimation q_Φ of the true posterior p_Θ, where Φ and Θ are the sets of all relevant underlying parameters for q (i.e., weights in the neural network) and p (i.e., parameters of the underlying true distribution). q_Φ consists of several conditional VAEs including an (optional) encoder q_Φ^L, for the observed (log-)library sizes of the samples allowing the library size to vary during model fitting.

The first decoder maps the latent spaces to fitted count matrices under an NB or ZINB distribution via three neural networks f_1, f_2, f_3 (Fig. 3C). Neural network f_1 estimates the fraction of reads for each transcript in each sample $c_{m,g}$ along with an inverse dispersion value θ. Neural network f_2 estimates the library size l_m of each sample m. θ and $\theta / l_m \cdot c_{m,g}$ serve as the shape and rate parameter, respectively, to a Gamma distribution to form counts $\omega_{m,g}$ which serves as the rate parameter to a Poisson distribution and together define a NB distribution with mean $\mu_{m,g}$ and dispersion θ (see the Supplementary Information for a more detailed explanation and also [45], Supplementary Note 3). If the user prefers to work with a zero-inflated model, a third neural network f_3 estimates the logit of the dropout rate $\pi_{m,g}$ for each transcript g in each sample m. $\pi_{m,g}$ is used as the parameter of a Bernoulli distribution that models whether a transcript is dropped out (therefore a zero). When combined with the NB distribution of c_{mg}, this results in a ZINB distribution.

Loss is computed as a (possibly weighted) combination of the Kullback-Leibler (KL) divergence and reconstruction error (log-likelihood of the NB or ZINB as appropriate). A more detailed description of the neural network with details on dropout, choice of activation layers, injection of correction variables, and optimization is described in the Supplementary Information.

Imputation of missing transcript counts

Several experiments were carried out to test the capacity of PREFFECT to self-learn or *impute* missing values using synthetic or pseudo-synthetic datasets. The collection D of positions (i, j) in the count matrix that were replaced by a zero to simulate dropout was recorded. The median relative error was computed between the true (hidden) value and the imputed value across all such locations (i, j):

$$MRE(X, \hat{X}) = median_{(i,j) \in D} \frac{|\widehat{X_{i,j}} - X_{i,j}|}{X_{i,j}},$$

where $\widehat{X_{i,j}}$ is the estimated value for this count and $X_{i,j}$ is the observed count.

Experiments to measure batch corrections

Several experiments were carried out to investigate the capacity of PREFFECT to adjust for batch effects. In the first experiment, we created a synthetic dataset with $M = 1000$ samples and $N = 900$ transcripts and simulated a simple batch effect as follows. First, a sample s was assigned to batch 0 with probability b. Otherwise, it was assigned to batch 1. Let $batch(s)$ denote its batch. Second, a frequency vector ω for the transcripts was generated using the stick-breaking algorithm [63]; ω was used to generate variates for all samples regardless of batch, and a single suitably large library size L was chosen for all samples. Transcript counts $C_{s,g}$ for sample s and transcript g were generated according to a hierarchical model as follows:

$$C_{s,g} \sim NB(\mathcal{N}(\mu_g, \sigma^2), \theta) + batch(s) \cdot \text{Bernoulli}(p) \cdot \mathcal{N}(\mu_B, \sigma_B^2),$$

where $\mu_g = L \cdot \omega_g$, $\sigma = 1$, p corresponds to the probability a transcript is subject to the batch effect, and μ_B and σ_B^2 are the parameters for a normal distribution describing the batch shift. Note that the final library sizes (from summing over all transcripts per sample) will tend to be larger for batch 1 samples.

Three distinct experiments were conducted as follows:

Experiment 1. Library size $L = 10^6$; $B = 100$; $\sigma_B^2 = 1$; $p = 1$ implying all transcripts in batch 1 received the adjustment. Samples were assigned to the two batches with equal probability.

Experiment 2. As for Experiment 1 but the probability p that an individual transcript in batch 1 would be subjected to the batch effect varied from 0.5 to 1.

Experiment 3. The goal of this experiment was to test whether important biological differences between the samples (here represented by samples belonging to two distinct subtypes) were not lost after adjusting for the batch effect. Similar to Experiment 2, but two distinct frequency vectors ω_α and ω_β were randomly generated via the stick-breaking algorithm representing two subtypes α and β respectively. A sample was randomly assigned to either subtype α or β with equal probability. Therefore, a transcript g in sample s, μ_g is equal to either $L \cdot \omega_\alpha$ or $L \cdot \omega_\beta$ depending on whether s was assigned to subtype α or β.

The *Jensen–Shannon divergence*, a symmetrized and smoothed version of the KL divergence, is used to measure the distance between frequency distributions the generative transcript frequencies ω and the estimated transcript frequencies $\hat{\omega}$. The JSD is defined as follows:

$$JSD(\omega \mid\mid \hat{\omega}) = \frac{1}{2} KL(\omega \mid\mid M) + \frac{1}{2} KL(\hat{\omega} \mid\mid M),$$

where $M = \frac{1}{2}(\omega + \hat{\omega})$ is the average, mixture distribution and $\mathrm{KL}(\cdot, \cdot)$ is the KL divergence. To measure the change in JSD after adjusting for the batch, we compute

$$\Delta\ JSD(\omega, \hat{\omega}, \bar{\omega}) = JSD(\omega, \bar{\omega}) - JSD(\omega, \hat{\omega}),$$

where $\bar{\omega}$ is the adjustment (via the latent representation) of $\hat{\omega}$.

Using the estimated transcript frequency vectors ω, the samples were mapped to two dimensions for visualization using UMAP.

Single-tissue PREFFECT model: sample-sample relationships.

Given the challenges of fRNA-seq data, we sought to incorporate additional information that could assist with the de-noising and imputation of count data. Towards this end, the *single tissue* generative model incorporates a sample-sample adjacency matrix $A : M \times M$ (Fig. 3A, purple). Two samples are adjacent if and only if they are deemed sufficiently similar. Often similarity is defined by Pearson correlation distance or another metric applied to a subset of transcripts, but other techniques could be used either directly with the count matrix X or some other independent datasets (not necessarily fRNA-seq), opening avenues for integrating multi-omics data, clinical information or other modalities. In the experiments related to clustering with the single model, edges were included in the graph if and only if the two samples were of the same PAM50 subtype.

In the single tissue generative model, the encoder processes the count matrix X, the adjacency matrix A alongside associated metadata K through a neural network to project the data into a latent space Z_A. The top layer of this neural network uses a graph attention network (GAT) [47, 52], which takes as input transcript counts for each sample in addition to the sample-sample adjacency list (Fig. S3D and Supplementary Information). The attention mechanism helps the model focus on the most informative dependencies while minimizing the effect of others. GAT mechanisms also facilitate hierarchical feature learning by aggregating information over multiple layers and at different levels of granularity. As before, the (log) library sizes ($log\ L$) can be allowed to vary during training. The encoder produces an estimation $q_\Phi(Z_A, Z_l | X, A, log\ L, K)$ of the true posterior distribution p_Θ, where Φ consists of all parameters of the neural network and Θ represents the true parameters underlying the data. Linear layers are applied after the GAT to form the final latent encoding Z_A of the graph and count matrix. The decoder reconstructs an estimate of the count matrix X and the adjacency matrix A (Fig. S3E, pink).

Investigating the contribution of the adjacency matrices

We sought to demonstrate that the inclusion of sample-sample adjacencies in the single tissue model increased performance over the simple model. Our explorations related to imputation, clustering and network reconstruction used the breast cancer datasets from the compendium. In particular, we needed to generate a (large) set of pseudo-synthetic samples using a transcript frequency vector ω_s specific to each subtype s. We began by preparing each dataset as described above and determined the PAM50 subtype of each sample. Next, we transformed the raw input count matrix X to a matrix X'' where

$$X''_{i,j} = \frac{X_{i,j}}{\sum_{j'} X_{i,j'}},$$

so that the resulting rows (samples) of X'' correspond to the frequency vector for all transcripts. Now for each subtype s, we compute a vector ω_s which is the average frequency of each transcript in all samples of the subtype s. ω_s is then used to generate sample-specific variations across each of the transcripts. For each subtype s, a chosen library size L, and transcript specific dispersion θ, we create N_s pseudo-synthetic samples where each such sample has a transcript count vector formed as follows:

$$\begin{array}{rcl} \mu_g & = & L \cdot \omega_s, \\ \omega_{i,g} & \sim & \Gamma(shape = \theta_g, rate = \theta_g/\mu_g), \\ X_{i,g} & \sim & Poisson(\omega_{i,g}), \end{array}$$

The relationship between μ_g, $\omega_{i,g}$, and $X_{i,g}$ is a standard way to derive variates according to a negative binomial distribution. Here $X_{i,g}$, the final count for transcript *g* in sample *i*, is distributed according to a *NB* distribution with parameters mean μ_g and dispersion θ_g. An adjacency matrix was formed as described above with edges placed between two samples if and only if they are of the same subtype.

Full PREFFECT model: integrating multiple matched tissues

The full model in PREFFECT generalizes the single model by accepting a set of tissue-specific count matrices $X^{(1)}, X^{(2)}, \ldots X^{(\mathcal{T})} \in \mathbb{Z}_{>0}^{M \times N}$ along with corresponding sample-sample adjacency networks $A^{(1)}, A^{(2)}, \ldots A^{(\mathcal{T})}$ for the $\mathcal{T}$ matched tissues profiled over a common set of *N* transcripts (Figs. 3D, S3F). The learning procedure fits the NB or ZINB models to each tissue via the GAT layers and adjacency matrices as in the single layer model, but additional layers in the network combine the $\mathcal{T}$ latent spaces into a single joint latent space. In this way, the related tissue types influence the fit and therefore the adjusted transcript count data.

Investigating performance of the full model

To investigate the performance and behavior of the PREFFECT full model, we require at least two matched transcript count matrices. For the purposes of this experiment, we turned to a pseudo-synthetic approach that exploits an existing dataset (Sunnybrook) in the fRNA-seq compendium with both tumor (first matrix) and matched stromal (second matrix) count data. In both cases, 100 transcripts across 1000 samples were generated using the parameters learnt from the real data for each tissue. Here sample *i* in the primary tumor matrix is matched with sample *i* in the stromal matrix.

For the primary tumor count matrix, we assigned 250 (of 1000) samples to each of four breast cancer subtypes. The first 50 transcripts correspond to PAM50 genes. NB counts were generated from parameters estimated for each subtype from breast cancer datasets in the fRNA-seq compendium. The second 50 transcripts are non-informative; the counts are simply variates from $\mathrm{NB}(100, 1)$ and so they should not impact clustering of the tumor samples This gives us a set of $M = 1000$ samples each assigned a tumor subtype across the 50 PAM50 transcripts and 50 non-informative transcripts.

To simulate stromal subtyping schemes, we partitioned the 1000 samples into 4 stromal subtypes (labelled $\alpha - \delta$) as follows:

- if the sample was assigned the basal subtype in the tumor samples, it randomly receives either stromal subtype α or β with equal probability;
- tumor HER2-enriched samples were assigned stromal subtype γ; and
- tumor luminal A and B samples were both assigned stromal subtypes δ.

In this (artificial) manner used for exemplary purposes, the stromal subtypes provide greater refinement of the tumor subtypes in one case (basal subtype), and the tumor subtypes provide greater refinement of the stromal subtype in one case (the δ subtype is fractured into luminal A and B subtypes).

A distinct transcript frequency vector was generated via a stick-breaking algorithm for each of the 4 stromal subtypes as described above for our investigations of batch corrections. The stromal count matrix consists of the 50 PAM50 transcripts and 50 stroma-specific transcripts. For the stroma, the PAM50 transcript counts are just random variates from $\mathrm{NB}(100, 1)$ so they are uninformative and should not affect clustering of the stroma samplings. The 50 stroma-specific transcript counts are generated using the corresponding frequency vectors $\omega_\alpha, \omega_\beta, \omega_\gamma, \omega_\delta$.

Training and hyper-parameter tuning

During the design and testing of PREFFECT, a large hyper-parameter search was conducted across the fRNA-seq compendium and pseudo-synthetic data. Our goal was to identify good default values for a wide range of parameters both with respect to specific values (e.g. default latent dimensionality) and architecture (e.g. type of activation function, number of linear layers). The ~ 25 parameters are detailed in Table S3 along with the default value. All of these parameters can be changed by the user, although it is often the case that only a few such parameters must be explored when fitting to a new dataset, primarily learning rate, epochs, batch size, latent dimension *r*, and weight of the KL divergence score for expression. PREFFECT expects distinct training, validation and test sets for learning a model with a recommended size ratio of 6:2:2. Only the training and validation datasets are used during model learning. The Supplementary Information provides a detailed description of the model and loss functions. The online software has a series of vignettes that aid the end-user with training models for new datasets; a vignette is depicted in Fig. S4.

Results

Table S1 describes the path of our analyses beginning with the characterization of fRNA-seq data, through the construction of a series of PREFFECT models with

increasing complexity, to the validation and exploration of PREFFECT on contemporary fRNA-seq datasets.

FFPE-derived RNA-seq expression is well-modeled by the negative binomial distribution

To characterize the distribution of transcript counts obtained from whole transcriptome fRNA-seq profiling, a compendium was constructed by selecting publicly available datasets which contain a considerable number of profiles (at least 20 samples; median=93) and for which the corresponding raw, non-normalized count data is available. The $N = 13$ sets of fRNA-seq samples vary in terms of tissue types, cell types, disease (or normal samples), age of samples, and technical variables (Table S2). The transcripts with the highest counts tend to be the same as those observed in bulk and scRNA-seq studies including *MALAT1, NEAT1, XIST* [64] and others listed in Fig. 1A. These extreme counts in the right tail induce a very high standard deviation with respect to the average count per transcript (Fig. 1B). For example, the mean count for *MALAT1* in the TMBC dataset [57] is ~ 4.4 million but the maximum observed value is 34 million. Mean trimming of the right-most 1% of extreme observations reduced the overall standard deviation by an order of magnitude in almost all datasets (Fig. 1A). We note, however, that not all of the extreme observations are artifacts that can be safely trimmed. Analysis below suggests that many high count transcripts occur in transcripts including ERBB2 (HER2) and samples where over-expression is expected. It should be stressed that even after trimming, the range of counts for many transcripts in fRNA-seq data remains very broad. For example, transcript counts for the *estrogen receptor 1 (ESR1)* in the Sunnybrook fRNA-seq dataset range from just 1 to more than 100,000.

Six different distributions were fit to the data using the AIC as a measure of fit. A significant majority of transcripts are best fit by the NB distribution for all but the normal tissue samples of GSE47462 (Fig. 1C). Because the various distributions differ in their number

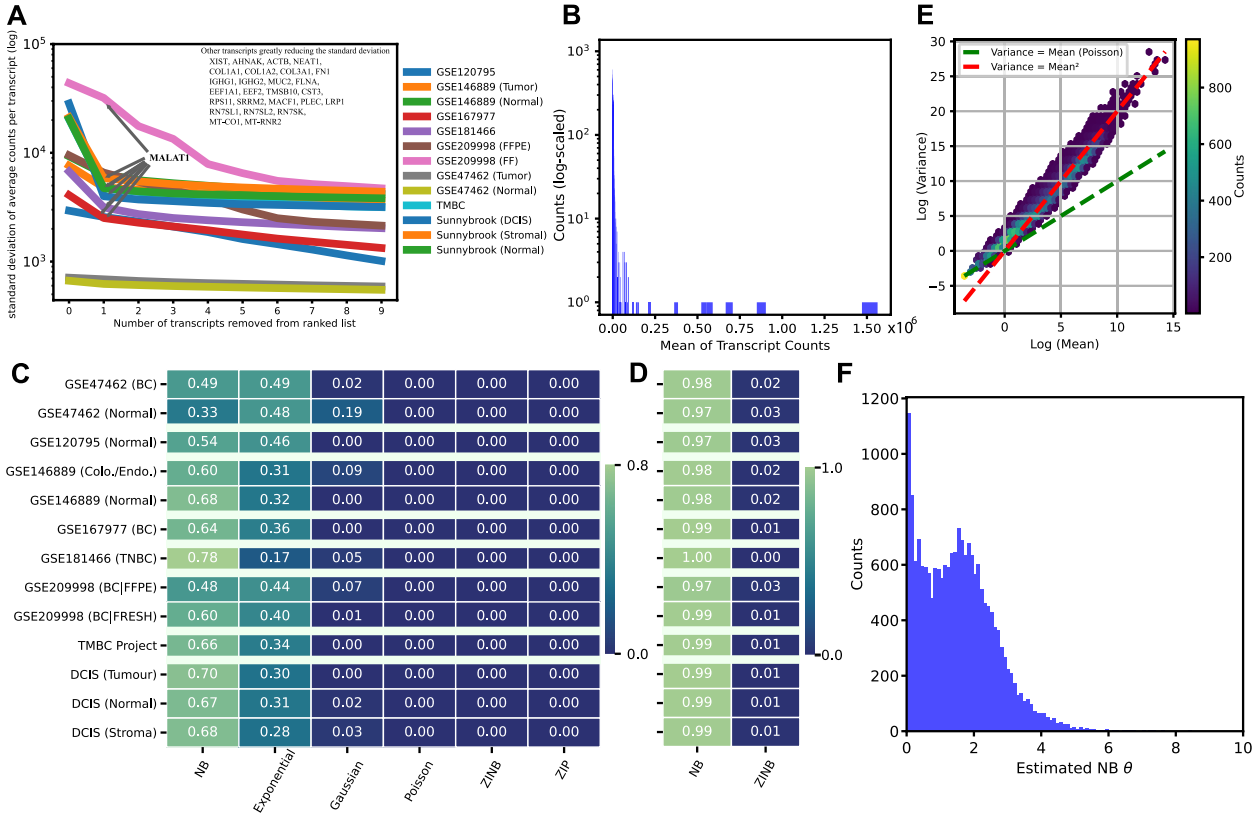

Fig. 1 Distributional properties of fRNA-seq data. **A** For each dataset in our compendium, transcripts were ordered from highest to lowest count. Depicted here is the number of highest transcripts removed from each dataset versus the reduction in the standard deviation. *MALAT1, NEAT1, XIST, COL1A1/2, MT-* mitochondrial and *RN-* ribosomal genes are highly and ubiquitously have many reads likely due to their large size. The standard deviation drops by an order of magnitude after the highest 5 transcripts are removed. **B** Histogram of the (log) mean count for transcripts in the *GSE209998* dataset. Note the extreme outliers $> 10^5$ counts. **C** For each dataset (row), the fraction of all transcripts with the best fit by distribution (column). The NB distribution has the highest frequency for all but *GSE47462*. **E** The vast majority of transcripts are better fit by the NB than the ZINB under the AIC criterion. **E** Hexagonal heatmap of the log mean versus the log variance of each transcript in *GSE209998*. The green line indicates the trend if the variance equaled the mean, while the red line indicates the trend for a quadratic mean-variance relationship. **F** Distribution for estimated dispersion θ with dataset *GSE47462*

of parameters and the AIC contains a weak penalty for model complexity, we repeated the analysis using the KS D statistic, which does not adjust for model size, but arrived at the same result (see Supplementary Information and Fig. S1A).

Figure 1C also suggests that transcripts often prefer the exponential distribution. However, we observed that this preference is highly dependent on the percentage used to trim outliers. That is, the preference for the exponential fit is strongly influenced by the number of highly expressed transcripts in the right tail (see Supplementary Information and Fig. S1B), further highlighting the fact that fRNA-seq data is prone to extreme measurements.

A significant fraction of transcripts have zero counts in all of the fRNA-seq datasets (ranges from 0.1 to 0.5 with median 0.46; Table S2). Zero-inflated extensions to distributions like the NB are often considered in such cases. In addition to the mean μ and dispersion θ of the NB distribution, the ZINB has a third parameter π that controls the probability of a so-called *dropout event* (a zero count for a transcript that is not a variate from the NB). However, Fig. 1D reports very little support for the ZINB. A direct comparison between the NB and ZINB reaffirms that observed zero counts are almost always well-modeled by the NB alone (Fig. 1D). Additional analysis in the Supplementary Information provides evidence that this is not due to model complexity. It is common in the context of NB-related distributions to express the variance as a function of the mean and the dispersion parameter θ as follows:

$$\sigma^2 = \mu + \theta \cdot \mu^2,$$

highlighting the fact that high values for θ induce a large variance in the NB distribution. This variance often appears to be sufficient to model the observed quantity of zero counts in fRNA-seq data, a conclusion which has also been reached for scRNA-seq data [41]. Figure 1E reaffirms the choice of an NB over a Poisson distribution, since the variance is observed to be greater than the mean.

A previous fRNA-seq transformation uses a probabilistic model where each transcript across all samples is decided to be modelled using either exclusively a (truncated) Gaussian or zero-inflated Poisson (ZIP) [32, 33]. Across our fRNA-seq compendium, their algorithm assigns only a small minority of all transcripts $< 17\%$ to the ZIP; the remainder are fit to a Gaussian. We observed very little support for the Gaussian, and found even less support for the ZIP (Fig. 1C, third and final column, respectively).

After fitting an $NB(\mu, \theta)$ to each transcript within each dataset, the overall mean μ is observed to be 710 but with an extremely large standard deviation, ranging from 666 to 43,713 (Table S2). We also observe a mean dispersion θ of 1.47 with a large relative standard deviation of approximately 1.5. Few transcripts have an estimated θ below 0.01 and few transcripts have an estimated θ above 5. An NB distribution with $\theta < 1$ is maximized at 0 with many near-zero counts, whereas larger θ are maximized strictly above 0. Many transcripts, including *ERBB2/HER2* and *ESR1*, have larger values for θ greater than 3 (Fig. 1F); these induce a flat, long NB distribution.

fRNA-seq data as a mixture of distinct technical and biological effects

It is well-established that technical variables often have a significant impact on distributional parameters in -omic profiling. This includes variables such as library size, batch number and library complexity (e.g., percent duplication and DV200). Across the fRNA-seq compendium, library size was highly dependent on batch number in all datasets where this information is available (one-way ANOVA, all $p < 0.001$). Moreover, both the location μ and scale θ parameters for the fitted NB distributions are almost always significantly different between batches (log-likelihood-based test of the change in fit between batch-dependent parameters). The percentage of duplicate reads also differed significantly between batches ($p < 0.01$ for all available datasets).

Biological variables (e.g., hormone receptor status, proliferative index, grade in cancer studies) should of course have a significant impact on the fRNA-seq profiles. To investigate how count distribution is affected, we focused on patient subtype across the breast cancer datasets within the compendium. It is well established that breast cancer samples can be partitioned into at least five distinct subtypes at the transcriptional level. Moreover, the expression of many genes is strongly subtype dependent. PAM50 is a commonly used classification tool that subtypes samples according to the counts of 50 specific transcripts for invasive [59] and *in situ* [65] lesions. Large differences in fitted NB parameters were observed when samples were stratified by the PAM50 subtype for the vast majority of the transcripts (Figs. 2 and S2). For example, the location parameter of the NB distribution for estrogen receptor 1 (*ESR1*) is markedly higher for luminal A (dark blue) and B (light blue) estrogen receptor positive subtypes in comparison to all remaining estrogen receptor negative subtypes; this is consistent with the original data represented (Fig. 2A). The fitted subtype NB distributions for Keratin 5 (*KRT5*) have the largest location parameter for normal-like lesions (green), the subtype where it is highest expressed according to the original PAM50 manuscript [59]. The NB fits for *ERBB2 (HER2)* and *GRB7*, two genes in the 17q12 amplicon characteristic of HER2 positive tumors (pink), have location (μ) and scale parameters (θ) well above zero only in the HER2

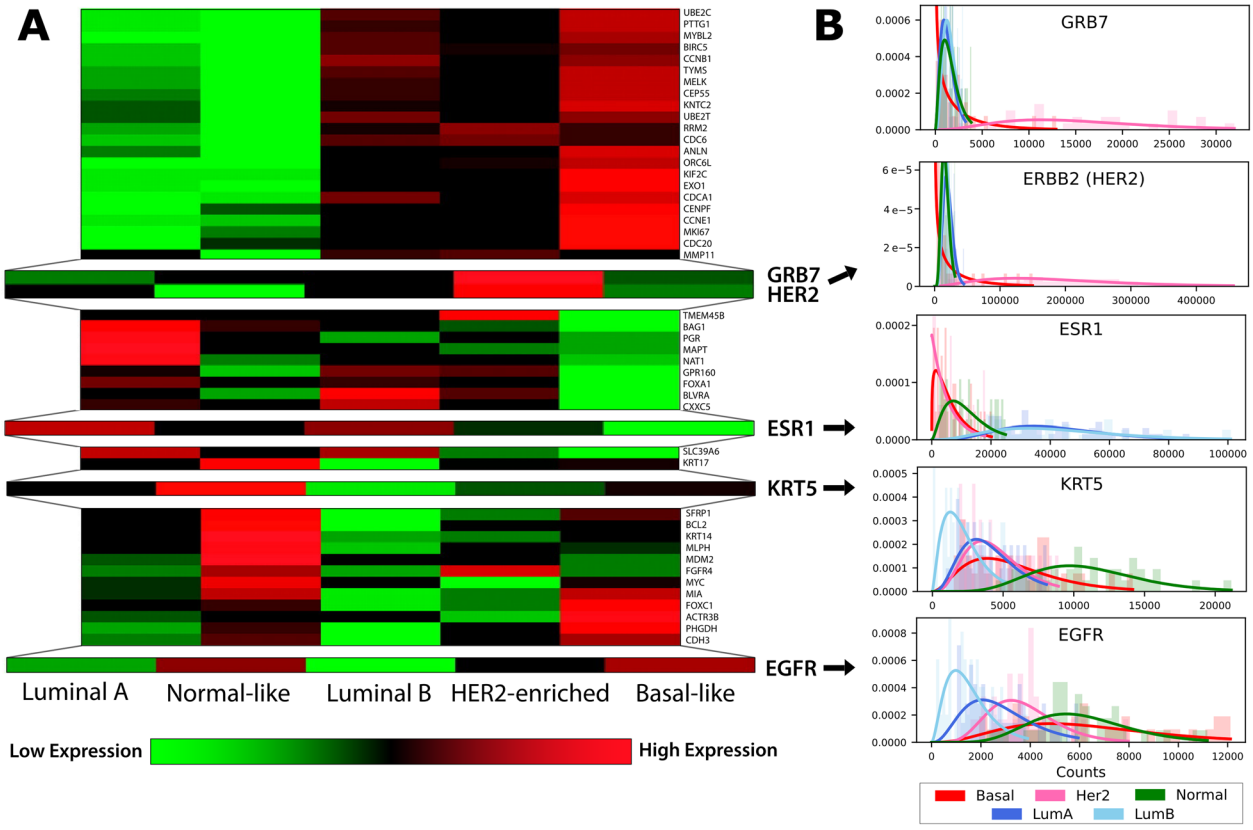

Fig. 2 The behaviour of the breast cancer subtype PAM50 transcripts in fRNA-seq data. **A** Heatmap modified from [59], depicting the expression of each of the PAM50 transcripts across the five breast cancer subtypes in the original data. Here, red and green depict over- and under-expression of the transcript respectively. **B** Histograms of transcript counts in the Sunnybrook DCIS tumor cohort for selected PAM50 genes (enlarged rows of the heatmap in **A**) colored by their subtype. A NB distribution was fit for each transcript in each subtype independently after median library size adjustment and trimming

subtype (Fig. e2B). Note also that *ERBB2 (HER2)* counts are at least one order of magnitude larger than all other transcripts, and there is a large variance in counts across the samples. Likewise, epidermal growth factor (*EGFR*), a well established marker of basal-like (red) and normal-like (green) tumors have similar NB fits in these two subtypes. The NB distribution here is flat, which corresponds to a high θ as discussed earlier.

It is clear from these examples and others (Fig. S2) that the distribution of a transcript is heavily influenced by biological effects, shown for subtype in this case. Because the approach from Yin et al. [32, 33] decides for each transcript over all samples whether it is modelled with a Gaussian or with a ZIP, it is not able in its current form to adapt to such effects. Instead, all transcript counts will be fit to a Gaussian, since their algorithm almost always assigns transcripts to this component, resulting in poor fits, especially lowly expressed genes. The mean and dispersion parameters of the NB distribution are far more flexible, providing good fits to a variety of distributional shapes (Fig. 4C).

We again do not see strong evidence for use of the ZINB over the NB distribution with any of the PAM50 genes when samples are stratified by subtype (Fig. S2). The need for the ZINB distribution would be clearly justified if we observed a large spike of zeros in cases where μ is well above 0. Instead, it appears that the NB distribution with a suitably high dispersion θ is sufficient in cases where μ is close to 0. We conclude that fRNA-seq datasets are well-modeled by the NB distribution albeit with large dispersion at times.

The simple PREFFECT model robustly estimates generative parameters when dropout rates are within observed ranges

Given the observations made above, PREFFECT was designed to model transcript count data by fitting it to NB or ZINB distributions, using observed transcript counts and conditioned by sample metadata. A conditional VAE is used to optimize the fits per transcript and sample while adjusting for technical and biological effects using a mathematical formulation (Fig. 3C) also used in scRNA-seq frameworks (e.g. scVI [45]), although the architecture of the model differs significantly through extensive train/validation/test based-learning with both

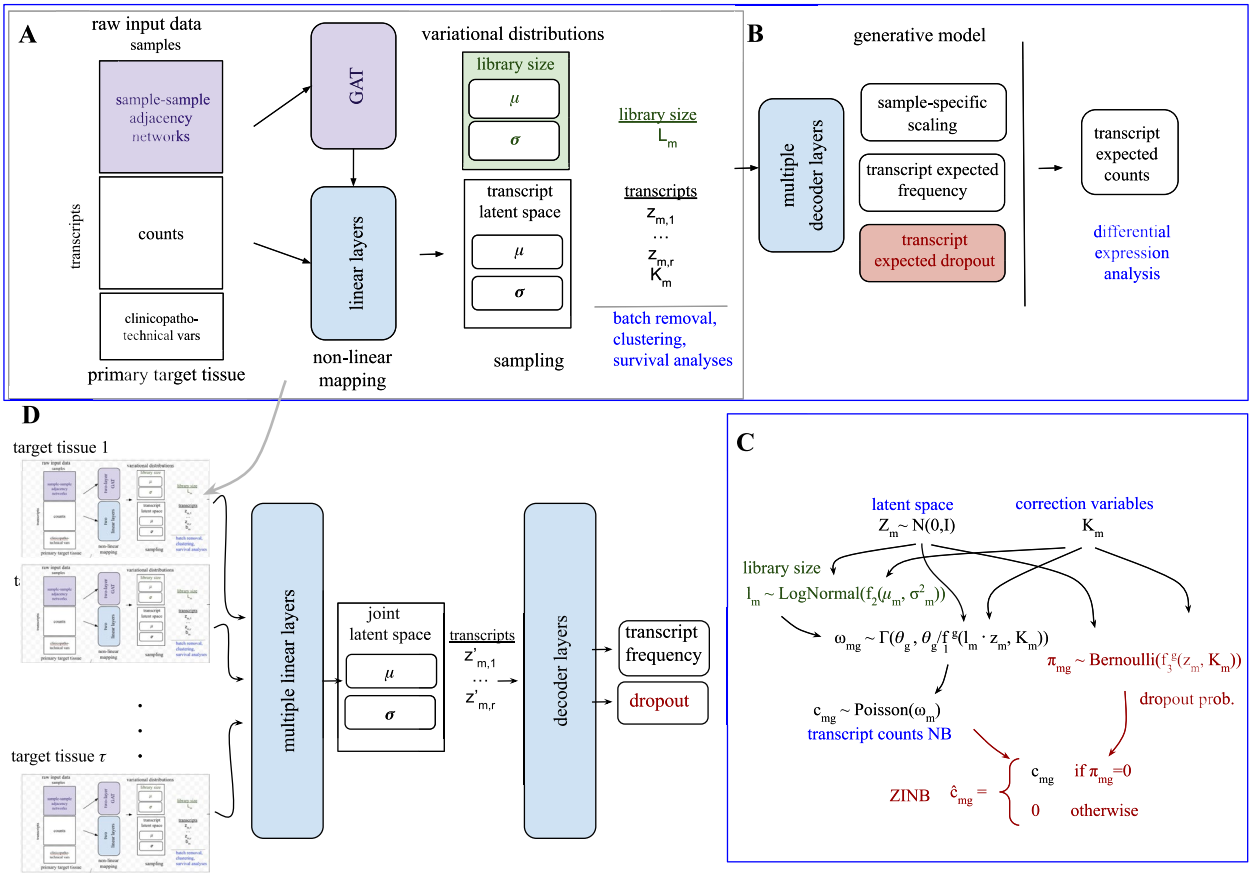

Fig. 3 Overview of PREFFECT. **A** The simple encoder consists of only the white boxes (count matrix and correction variables). The single encoder integrates the sample-sample adjacency graph with attention mechanisms (purple boxes). **B** The single decoder extends the simple decoder with multiple decoder layers that allow integration of the adjacency information. If a ZINB is the desired distribution, the decoder also estimates a dropout rate π denoted in red. **C** The relationships between all variables from the underlying statistical model. **D** The full model combines one single encoder model for each available tissue

real and pseudo-synthetic fRNA-seq data (Figs. 3A, B, S3, Methods, Table S3).

Using the distributional parameters from observed fRNA-seq datasets, we generated so-called *pseudo-synthetic* samples where we know the ground-truth counts and overall distributions. This was repeated across a range of values for both the μ and θ NB parameters which capture the behavior of the vast majority ($> 95\%$) of transcripts in the fRNA-seq compendium (see Methods). The pseudo-synthetic datasets allows us to measure the capacity of the simple PREFFECT model to recover generative parameters across a broad range of values. Figure 4 shows that performance is near perfect for both μ (panel (A) and θ (B) everywhere except when θ is very small (0.01). Such small values correspond to NB distributions with almost all zeros (top of panel *C*). In total, 99% of all transcripts have a θ larger than 0.01. Although the pseudo-synthetic data was generated using an NB distribution, the ZINB still accurately assessed the parameters μ and θ (second column of panel A and B).

The elevated number of zero counts for transcripts in fRNA-seq data motivated a study of how well PREFFECT can impute missing values. We used a simple self-learning approach to imputation where PREFFECT replaces masked values with the adjusted expected value from the estimated distribution. To explore this, we again used pseudo-synthetic data generated with an NB distribution as described above. However, now each transcript is subjected to dropout with a randomly assigned rate $\pi \in U(0 \ldots 0.8)$, producing a ZINB distribution with known dropout locations. Not surprisingly, the performance depicted in Fig. 4D–E suggests the quality of the fits inferred by PREFFECT are overall poorer than simulations without dropout when *NB* is used. However, when ZINB is used, the performance remains high especially for larger θ values, and decreases only for low θ, likely because the presence of many endogenous zeros (zeros not caused by dropout) leads to an inflation in the estimate of $\hat{\pi}$, consistent with the reduction in accuracy observed at the top of panel F.

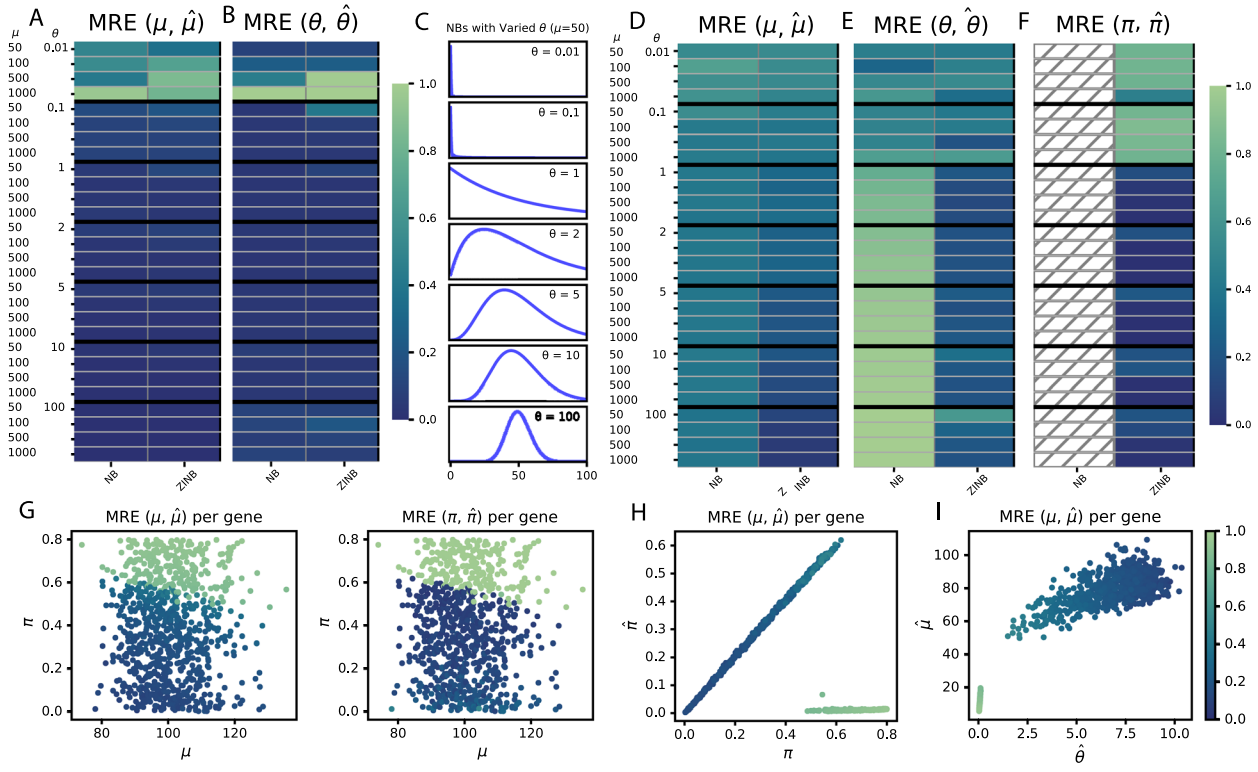

Fig. 4 The ability of PREFFECT to recuperate generative parameters. NB counts were generated for $N = 1000$ transcripts across $M = 1000$ samples across a range of parameterizations for μ and θ. PREFFECT was then used to infer parameter levels under either an NB (left column) and ZINB (right column) model. Colors in the heatmap correspond to the mean relative error (MRE) between the generative parameters μ (**A**) and θ (**B**) versus their respective estimates $\hat{\mu}$ and $\hat{\theta}$. The MRE remains very low for all parameters except when θ is very small. **C.** Examples of the effect of θ on the NB distributions. When θ is very small at 0.01, many transcripts have a zero count. **D, E, F** The ability of PREFFECT to recuperate generative parameters when challenged with dropouts. The same synthetic generative methods from panels **A** and **B** are repeated but now each transcript was subjected to random dropout (from 0 to 80% of all samples are set to 0). PREFFECT was used to infer parameters. As expected, the ZINB model is near universally better than the NB model, where μ (**D**), θ (**E**) and π (**F**) have low MRE except for small dispersion θ levels. **G, H, I** The performance of estimating ZINB parameters μ and θ with random amounts of dropout where $\theta = 10$. **G**: Color is proportional to the MRE of the masked positions for each transcript (point) plotted according to the generative parameters μ and π. **H**: Color represents the MRE relative to π and $\hat{\pi}$; and **I**: MRE relative to the generative $\hat{\mu}$ and $\hat{\theta}$

To examine the impact of the dropout rate π in parameter estimation, we constructed a second dataset designed in a manner that it would likely contain very few endogenous zeros by setting μ and θ appropriately (see bottom of Fig. 4C). Therefore, when we mask values in the generated count matrix using a random dropout rate $\pi \in U(0 \,..\, 0.8)$ for each transcript, the estimate $\hat{\pi}$ should be very close to π, since the vast majority of zeros are truly due to dropout. Figure 4G–I show that PREFFECT estimates μ and π well for $\pi < 0.6$ but degenerates for higher dropout rates. None of the datasets in our compendium had a dropout rate ≥ 0.55.

The simple model can accurately adjust for batch effects

Generative models can be used to hypothesize how a dataset might change as specific effector variables are modulated. For example, in large-scale projects, samples are prepared and profiled in batches, and this batch variable can systematically affect a count matrix. It is often necessary to adjust the batches to remove such effects before downstream analyses. We explored the capacity of PREFFECT to identify and ablate batch effects.

In the first experiment, counts were generated for all samples using a family of NB distributions with location parameters determined by a single underlying transcript frequency vector ω. The samples were then randomly assigned to batch 0 or 1, but only counts for transcripts in batch 1 were systematically increased to simulate the batch effect (see Methods). As expected, after training, samples clearly cluster by batch number when the frequency vectors are computed from the observed count matrix (UMAP, Fig. 5A). However, by shifting samples in batch 1 towards batch 0 in the latent space, the resultant adjusted count matrices no longer cluster by their batch (Fig. 5B). Figure 5C confirms that the frequency vectors computed from the simulation differ between batch 0 and 1, as expected. If PREFFECT successfully fits a good model, the difference in values between batch 0 in panel C and panel D will be marginal. The same statement holds for batch 1 between panels C and D. Lastly,

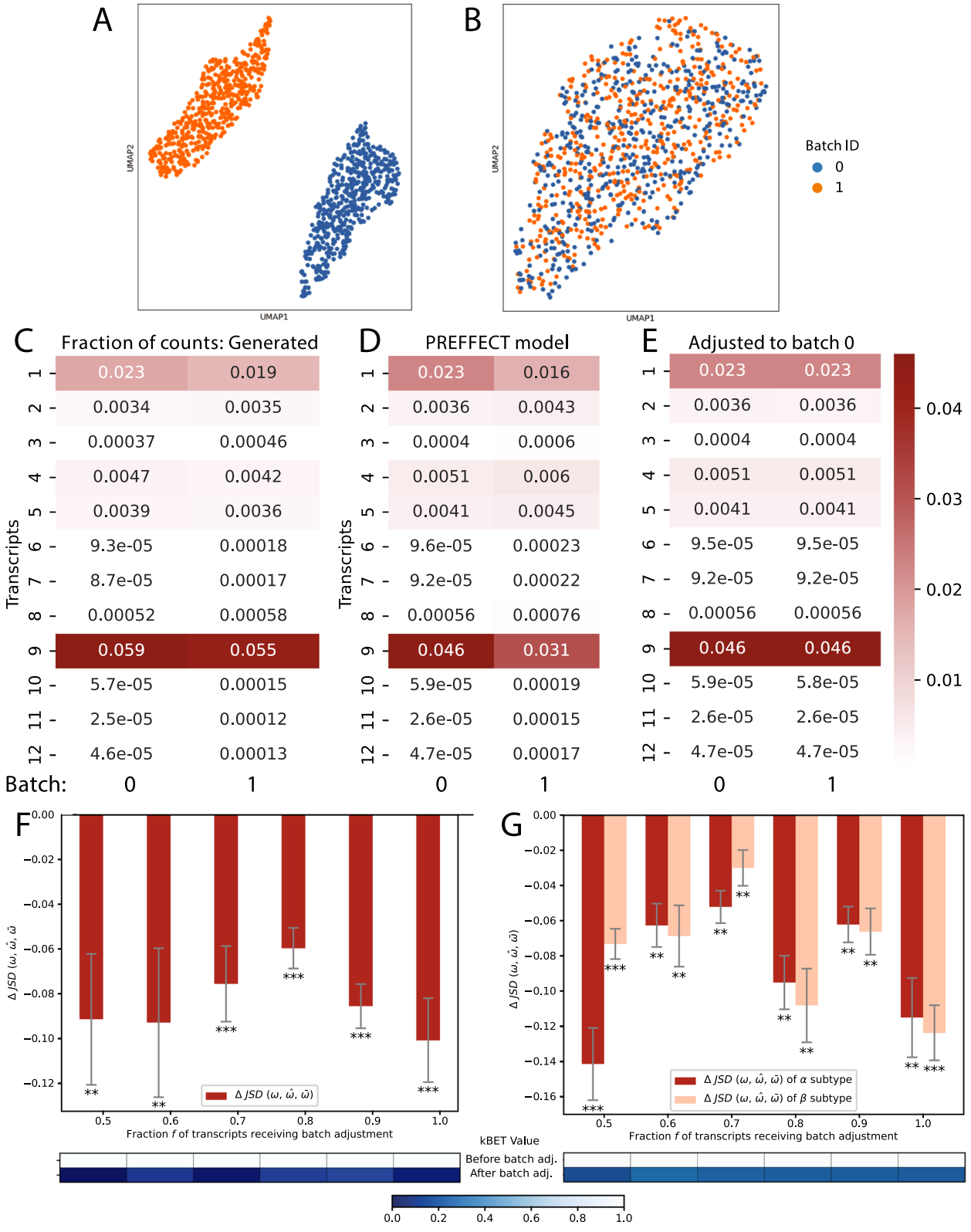

Fig. 5 (See legend on next page.)

(See figure on previous page.)

Fig. 5 Adjusting for batch effects. A PREFFECT model was derived using synthetic data with a simulated batch effect on all transcripts with a randomly chosen subset of samples. **A** UMAP embedding using the frequency vector ω_s for each sample *s* obtained from the raw count data. **B** UMAP embedding using the estimated frequency vectors $\hat{\omega}_s$ after adjusting the latent space of batch 1 to batch 0. **C** The average fraction of all counts for each transcript between batches in the raw synthetic data. **D** After training, these same fractions are retained when no adjustment is carried out. **E.** During inference, all transcripts were adjusted to batch 0. **F, G** Correction with respect to fractional batch effect. Synthetic datasets were generated where counts for a fraction *f* of the transcripts ($f \in \{0.5, 0.6, \ldots, 1\}$) in batch 1 were subjected to the effect. **F** The difference in the Jensen-Shannon divergence (JSD) was computed between the true generative transcript frequencies ω and $\hat{\omega}$, and between ω and the estimated frequencies after adjusting for the batch effect $\bar{\omega}$. For all fractions *f*, the $\Delta\mathrm{JSD}$ is negative implying that the adjustment has shifted the estimated frequencies closer to the true generative distributions. $*, **, ***$ denote $p < 0.05, 0.01, 0.001$ resp. derived from a *t*-test of whether the ΔJSD always improved. The *k*-BET measure of cluster mixing was always 1 (no mixing) before batch adjustment; blue color bars show *k*-BET after batch adjustment. **G** Similar to F but here samples were randomly assigned to either the α or β subtype with distinct transcript frequency vectors ω_α and ω_β respectively. Heatmaps below panels F and G visualize kBET across UMAP-defined sample clusters pre- and post-adjustment

by shifting batch 1 samples towards batch 0 in the latent space, batches 0 and 1 will have nearly identical frequency vectors, as observed in panel E, indicating a successful ablation of the batch effect.

The second experiment tests the ability of PREFFECT to identify and adjust for a batch effect at different levels of pervasiveness. Here, a series of PREFFECT models were fit to a count matrix similar to the first experiment, but only a fraction p of the transcripts in batch 1 received the batch adjustment. We computed the difference between two similarities: (ii) the similarity between the generative transcript frequencies ω and the estimated frequencies $\hat{\omega}$, and (i) the similarity between ω and $\hat{\omega}$ after adjusting for the batch effect with the latent space, denoted $\bar{\omega}$. For all values of p, the batch adjustment $\bar{\omega}$ is more similar to the true generative ω (Fig. 5F).

The third experiment was designed to ensure that important biological variation is retained after batch adjustment. Figure 5G extends the previous exploration to investigate cases where the samples differ by both their batch and their subtype (representative of a biological effect). To simulate this, samples were randomly assigned (with equal probability) one of two subtypes α and β each with a distinct frequency vector $\omega_\alpha, \omega_\beta$ respectively. Again, regardless of the pervasiveness p of the batch effect, the adjustment increases the amount of mixing between the two classes. This can be seen by the decrease in the kBET scores. At the same time we observe that the adjustment increases the degree of similarity with the two true generative frequencies. This means that the important biological differences are retrained since the frequency of the transcript counts tend towards the ground-truth generative vectors.

Lastly, we note that although the batch adjustment capacity of PREFFECT has advantages, PREFFECT transformed count data can also be used with other well-established tools (e.g. ComBat-seq [66]).

The single tissue PREFFECT model improves sample clustering

We sought to integrate additional information that could assist with the de-noising and imputation of the count data. Toward this end, the *single tissue* generative model extends the simple PREFFECT model by incorporating a sample-sample network (Fig. 3A, purple). This is achieved using so-called graph attention network layers, which are powerful neural network components that assist the artificial neural network to focus attention on the most informative components of the learning set during training. In our experiments here, two samples are adjacent if and only if they are deemed sufficiently similar. The exact notion of similarity can vary, providing a convenient means to integrate complementary types and modes of data.

We explored how the inclusion of the sample-sample network can improve the performance of downstream tasks, specifically sample clustering. Since available fRNA-seq datasets are limited in size, we generated a pseudo-synthetic dataset consisting of 200 samples for each of the 5 breast cancer subtypes (see Methods). Briefly, we estimated frequency vectors for PAM50 transcripts for each of the five subtypes depicted in Fig. 6A, generated a large set of pseudo-synthetic samples for each subtype, and ensured that the resultant samples had similar patterns of expression as the original PAM50 study (Fig. 2A). Two patient samples were made adjacent if and only they had the same subtype.

Not surprisingly, the single-tissue model trained with the network was able to recuperate the subtype-specific transcript frequency vectors ω (average $\mathrm{JSD}(\omega, \hat{\omega}) = 0.016 \pm 0.017$; Fig. 6B). Moreover, the UMAP produces five distinct clusters that nearly perfectly separate samples by subtype (panel F; Silhouette statistic 0.85). To test the contribution of the adjacency network, we repeated the training process but this time removed a fraction of all edges. Figure 6C depicts the estimated frequency vectors when 80% of all edges were removed, leaving only 20% of the edges between samples of the same subtype. There is a noticeable decrease in the model's capacity to recapitulate the generative transcript frequencies ω (average $\mathrm{JSD}(\omega, \hat{\omega}) = 0.041 \pm 0.029$). This is reflected in the associated UMAP where each "snake-like" cluster contains samples with different subtypes (panel G; -0.07 Silhouette). The last experiment instead

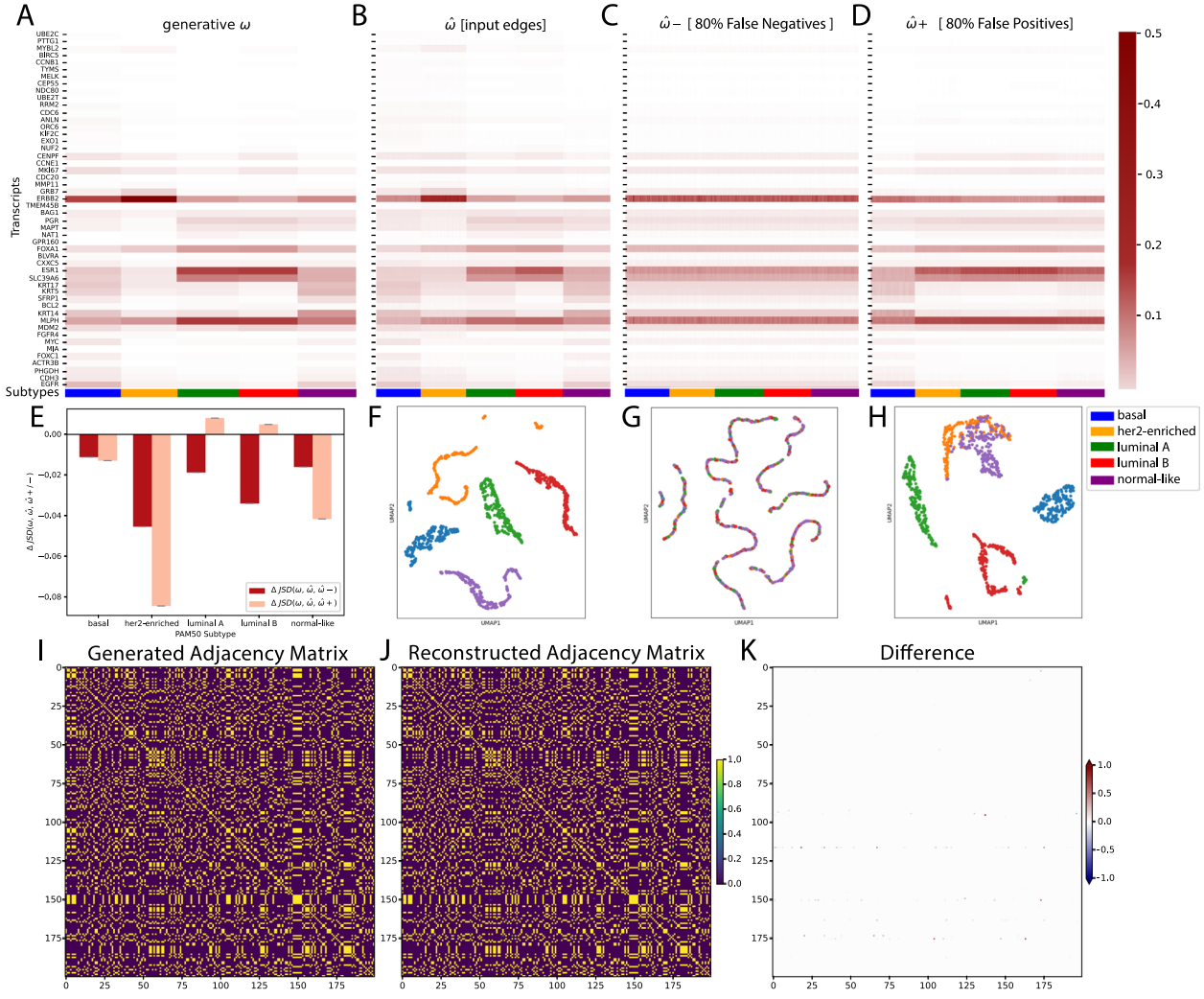

Fig. 6 The adjacency information assists in down-stream applications such as sample clustering. A pseudo-synthetic count matrix was constructed using NB distributions for the PAM50 transcripts derived from breast cancer fRNA-seq datasets. **A** A heatmap of frequency vectors ω for each of the 5 PAM50 breast cancer subtypes. **B** A heatmap inferred $\hat{\omega}$ of a PREFFECT model built using an informative sample-sample edge matrix. The UMAP is also clustered from the $\hat{\omega}$. **C** The $\hat{\omega}-$ when 80% of edges of the sample-sample adjacency matrix were randomly deactivated. **D** The $\hat{\omega}+$ when 80% of edges of the sample-sample adjacency matrix were randomly activated. **E** We compare the distributions of $\hat{\omega}$ from **B-D** to the generative ω (**A**) of each subtype via JSD. We plot $\Delta\ JSD$ of $\hat{\omega}$ (**B**) to either $\hat{\omega}-$ or $\hat{\omega}+$ (**C–D**) across each PAM50 subtype. **F-H.** UMAP clustering of $\hat{\omega}$ from **B-D**, respectively. **I.** The adjacency matrix for a random subset ($M = 200$) of the samples. Here yellow corresponds to the existence of an edge indicating both samples have the same subtype (probability of 1) and black corresponds to no edge (probability of 0). **J.** The reconstructed adjacency matrix. **K.** Differences between the generated and reconstructed adjacency matrices

introduces false positives into the adjacency matrix prior to training. Again, the capacity of the model to recapitulate the ω vectors is reduced but remains significant (average $\text{JSD}(\omega, \hat{\omega}) = 0.042 \pm 0.052$; 0.63 Silhouette; Fig. 6D) and the resolution of the UMAP has decreased with some HER2-enriched samples clustering with normal-like samples, and some confusion between luminal A and B (panel H). The reconstructed networks, which are allowed to evolve via distinct components of the VAE (Supplementary Information), are nearly indistinguishable from the original input graphs (Fig. 6I–K). These experiments show that PREFFECT is able to take advantage of the network information, which in this case allows the learner to focus attention on samples with the same subtype and therefore transcripts with similar count levels.

Full PREFFECT model: integrating multiple matched tissues

In many genomic-based clinical studies, matched fRNA-seq data is also available for related tissues or conditions in addition to the primary target tissue. For example, in disease studies, often both the affected and matched healthy/normal tissue from a patient is profiled. In cancer, the profiles of an index lesion can be complemented by profiles of their match normal tissue, the tumor microenvironment, and metastatic sites. The inclusion

of multiple matched tissues can improve performance, especially when there is significant dropout of transcript counts. Figure 7 provides an intuition of how this information is "borrowed" across profiles. In addition to the count matrices, the adjacency matrices are also updated during training to find the best fit possible (Fig. 3D and Supplementary Information).

To investigate the benefits of multiple tissue analysis, we turn once again to breast cancer as an example, since we have access to a dataset with matched primary tumor and stroma profiles and we know the ground truth subtype assignments for the tumor samples. It is well-established that breast cancer samples have both a primary tumor subtype [59, 67] and a tumor stroma subtype (e.g., [68]). The tumor and stroma subtyping schemes are distinct and the relationship between them appears to be complex and is still not fully understood. Our goal is to show that PREFFECT can use information from the stromal samples to form better sample clusters in the primary tumor, and vice versa.

To explore this, we generated pseudo-synthetic count matrices for both tissues as follows (see Methods for more details). Starting with the primary tumor matrix, samples were randomly assigned to four breast cancer PAM50 subtypes; in each sample 50 transcripts correspond to PAM50 and the rest with random counts. Next, for the stroma, each sample was assigned a subtype: basal samples were randomly assigned either stromal subtype α or β, HER2-enriched samples were assigned stromal subtype γ, and luminal A and B samples were assigned stromal subtype δ. A distinct frequency vector was generated for each stromal subtype and used to generate counts for the 50 stromal genes. Random values were assigned to the PAM50 transcripts in the secondary tissue dataset (that is, the PAM50 transcripts are not informative in the stromal samples). In this two-tissue scenario, we would expect that samples will cluster according to the four tumor subtypes when the tumor count matrix is used, but α and β samples would not separate. Conversely, we expect that the samples will cluster by the four stromal subtypes when the stromal count matrix is used, but the luminal A and B samples would not separate.

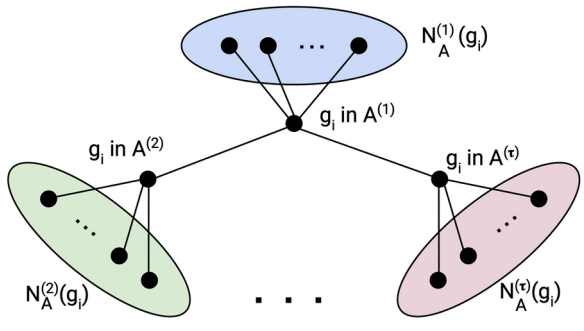

Fig. 7 Neighborhood information is encoded in the adjacency networks. The target transcript g_i in the primary tissue (blue) is influenced by the counts from its neighbouring samples in the primary tissue, but also by the expression of g_i in the other available tissues 2 (green) $\ldots \tau$ (pink), which are in turn influenced by their neighbours in the sample-sample network

Figure 8A confirms this hypothesis regarding separation of the luminal A and B subtypes when using the tumor counts. Interestingly, in the stroma-related Fig. 8B, we also observe the luminal A and B samples separated. We hypothesize that this unexpected result is due to backpropagation during model training, which transfers information from the model of tumor counts to the model of stromal counts. Regardless, when the combined latent space is used to cluster the samples, both the tumor luminal A and B subtypes and the stromal α and β subtypes are separated (Fig. 8C). This shows that the underlying artificial neural network is using information from both tissues when deciding the relationship between samples from both tissues.

PREFFECT models with contemporary fRNA-seq datasets

We examined the capacity of PREFFECT to fit good models with available fRNA-seq datasets, and tested whether the resultant models aided in downstream analysis, specifically sample clustering. Simple models were built for each dataset in the compendium, but our analysis below focuses once again on the six breast cancer-related datasets to investigate performance. Here, models were restricted to $N = 776$ genes from the well-studied pan-breast cancer BC360 panel (NanoString Inc.), since we can expect that these transcripts will vary significantly across the datasets. Initially, hierarchical clustering was applied to the data instead of applying PREFFECT (log-transform mean trimming 1% with variance stabilizing transformation). We observe a broad range of count values with many zero counts (represented by white). Although clusters are enriched for same-subtype samples (especially basal), many subtypes (especially luminal B) are diffuse across the clustering (Fig. 9A for dataset *GSE167977*). When the inferred transcript frequencies $\hat{\omega}$ are used instead, the expression becomes more polarized away from 0, presumably due to imputation of missing values. Panel B depicts the inferred transcript frequencies $\hat{\omega}$ using the same transcript and sample ordering as panel A for comparison purposes. When the samples and transcripts are re-clustered using the inferred transcript frequencies, we see much more homogeneous clusters for every subtype with the exception of the basal subtype which was already homogeneous (panel C). The adjusted rand index (ARI) increases from 0.32 for enrichment of subtypes in the sample clusters of panel A to 0.61 in panel C. Figure S5 depicts the hierarchical clustering of samples and subtype assignment analogous to panel A using only a VST (without PREFFECT). It has a slightly

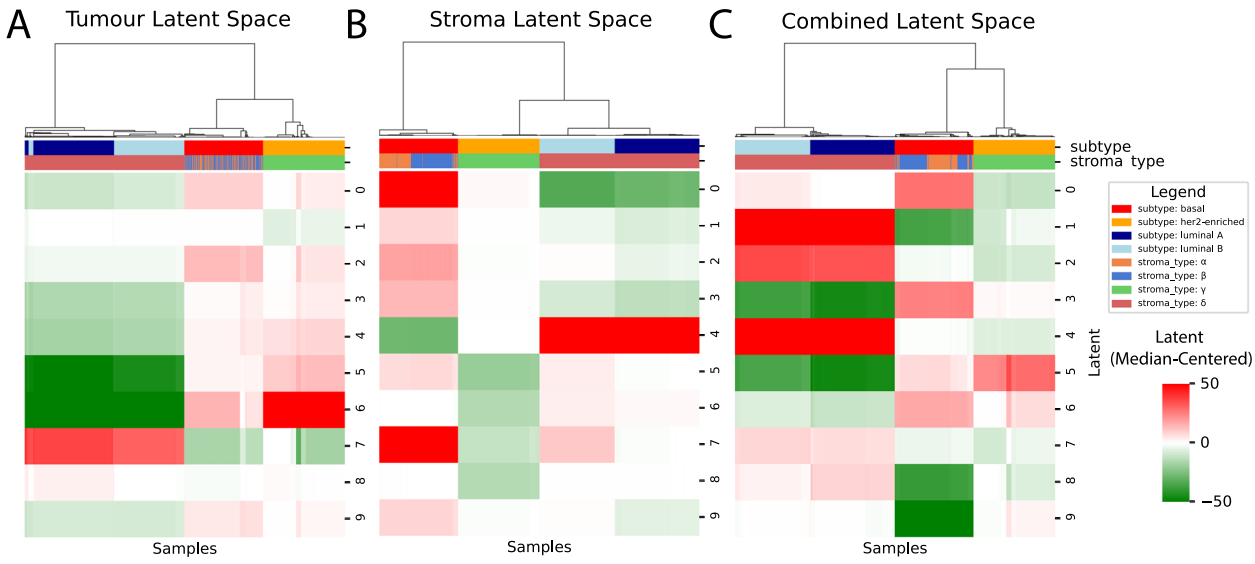

Fig. 8 Multi-tissue PREFFECT models can be developed using information from both primary and secondary matched tissues. To test the full model, we developed both a pseudo-synthetic breast cancer tumor count matrix (where PAM50 transcript counts follow subtype-specific distributions) and a secondary synthetic stromal count matrix (where a second set of 50 transcripts had counts separating them into four stromal subtypes $\alpha, \beta, \gamma, \delta$). Hierarchical clustering was applied to the resultant latent spaces **A** The latent space of the primary tumor tissue, which clusters according to PAM50 subtype but not stromal subtypes. **B** The latent space of the secondary stromal tissue, which clusters according to stromal subtype but also surprisingly separates luminal A from B in the tumor, likely due to information transfer during the learning procedure. **C** The combined latent space, which clusters the samples by the cross-product of the two subtyping schemes

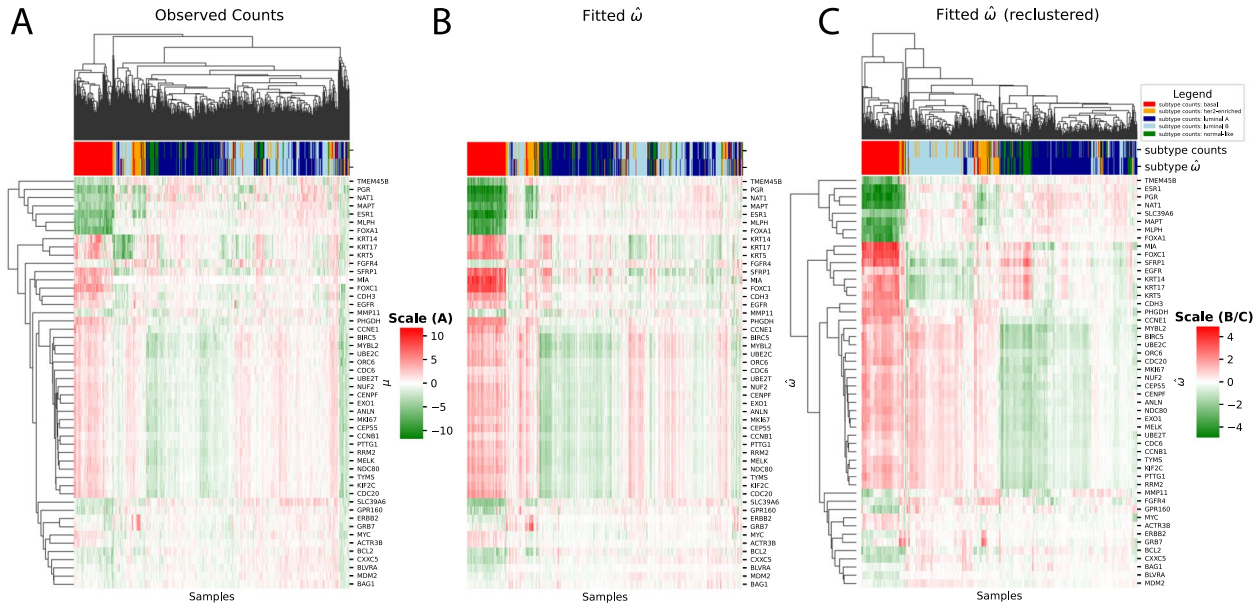

Fig. 9 PREFFECT models improve the quality of downstream patient clustering. A simple PREFFECT model was fitted using the BC360 (NanoString Inc.) panel from dataset *GSE167977*. **A** Hierarchical clustering was performed using the PAM50 transcripts contained in the BC360 panel. **B** The same sample and transcript clustering is re-drawn from panel *A* but instead here color corresponds to the estimated transcript frequencies $\hat{\omega}$. **C** Using the estimated frequencies $\hat{\omega}$, the samples and transcripts are re-clustered, resulting in sample clusters which are more homogeneous, consisting of a single subtype. In all panels, PAM50 subtypes were inferred from observed counts (top) and also from the inferred frequencies $\hat{\mu}$ (bottom)

poorer performance than the non-PREFFECT clustering of panel *A* with an ARI of 0.28.

Discussion

FFPE material is an important but under-utilized resource of well-preserved samples for both human health and disease. The profiles obtained from applying next generation RNA sequencing to FFPE material are noisy, prone to extreme measurements and contain a high zero count for transcripts (nearly one half at 0.45). The problematic nature of fRNA-seq underscores the importance of using the appropriate normalization and transformation as a first step in any analysis. Most studies to date have relied on techniques from bulk RNA-seq which often focus on fresh frozen tissues, cell lines or other forms of largely intact material. However, variance stabilization transformations are inoptimal for profiles with elevated zero counts. A typical scRNA-seq study, which shares some of the same challenges as fRNA-seq, has the luxury of orders of magnitude more cells than the number of samples in most fRNA-seq studies and it has a more restricted transcript count range.

Our analysis of the fRNA-seq compendium suggests that the vast majority of transcript counts are well-modeled by an NB distribution. This is consistent with other types of RNA-seq data, although both the mean μ and dispersion θ parameters vary considerably compared to other types of RNA-seq data including bulk and single-cell profiling. Although there is very considerable dropout, the NB is still able to model transcript counts well and we observed little support for use of the zero inflated extension (ZINB) with additional dropout parameter π. The NB distribution is well-established in expression profiling and serves central roles in many downstream applications including differential expression (e.g. DEseq2). We do not observe support for a previous effort (MIXnorm) which assumes that each transcript follows either a zero-inflated Poisson or a truncated Gaussian.

We focused here on breast cancer datasets and breast cancer subtype in our analyses. This restricted focus allowed us to comment on the performance of PREFFECT, since the behavior of many transcripts central to determining the subtype of a tumor is extremely well-characterized including the 50 transcripts of PAM50. In this sense, breast cancer subtype provides us with a gold-standard or "ground truth" to judge improvements in downstream applications post-PREFFECT.

Generative models allow observed data to be decomposed or "factorized" by such variables, whether they are known or unknown. PREFFECT is a series of generative models based on conditional VAEs to impute and factorize observed transcript count data to de-noise and adjust for both technical and biological variation. Unlike a previous fRNA-seq model [32, 33], PREFFECT has the capacity to modulate the distribution parameters in response to the state of biological variables such as patient subtype, an important capacity given the observed subtype-specific behavior of many transcripts. PREFFECT offers a number of alternatives to adjust for batch effect, and the PREFFECT adjusted count matrices can be easily used with other batch correction tools such as ComBat-seq [66].

We showed that PREFFECT can accurately infer generative parameters and accurately impute missing values for the range of values observed in the real data. Although PREFFECT performed well when evaluated on a test dataset with samples not seen during training, broader experimentation beyond our current compendium is still required. Imputation can be problematic, introducing for example bias, false correlations, and causing p-value inflation [69]. In short, imputation and other features facilitated by generative modelling are powerful tools for discovery in clinical FFPE cohorts, but additional due diligence is necessary if clinical diagnostics were to directly rely on the inferred information.

This single tissue model uses graph attention networks (GATs) to assist the learner to attend to the most influential neighbors (samples) from which infer distributional properties. We observed how such imputation can lead to better patient subtyping with breast cancer datasets. In general, the attention networks can be designed to integrate diverse types and modes of data into analysis. The full model allows for multiple matched tissues from the same patient sample to be integrated. To the best of our knowledge, this is the first generative tool to incorporate multiple patient-matched tissues and graph attention.

The vast majority of available fRNA-seq datasets currently are of moderate size with a median of 93 samples in our compendium, a value that is four orders of magnitude lower than some datasets available for scRNA-seq. Generative approaches such as PREFFECT, which provide a means to ablate nuisance technical parameters and better capture true biological signal, would certainly benefit from larger fRNA-seq datasets, given the degree of variability and extreme measurements, especially in contexts such as cancer where we know that samples are affected by strong transcriptional programs (e.g., estrogen receptor status in breast cancer and other subtype-related programs).

The undersized nature of current fRNA-seq datasets may partially explain why FFPE-based studies remain very challenging. Although PREFFECT is able to fit models to existing datasets, training required multiple runs with different parameter settings and we often had to make use of pseudo-synthetic extensions of these datasets. This can introduce subtle biases into studies. We conjecture that PREFFECT would benefit greatly from much larger sample sizes and result in more accurate

downstream analyses (e.g. differential expression, survival analyses, clustering).

Conclusions

RNA extracted from FFPE materials suffers from degradation, fragmentation, and chemical modifications that pose significant challenges for molecular analyses. Transformation and normalization of raw transcriptional data is a critical step that affects all types of downstream analyses needed for biomarker discovery and molecular characterizations. We developed PREFFECT to characterize and de-noise fRNA-seq data to enable more precise downstream analyses. such as differential expression, survival analysis, and clustering. PREFFECT performance was shown to be improved when information from multiple tissues and associated samples are leveraged to inform the graph attention mechanisms. PREFFECT is available as open source software and can be easily modified and extended. Our hope is that it serves as a central point for the community to reason about large-scale fRNA-seq studies.

Abbreviations

AIC	Akaike information criterion
ARI	Adjusted Rand Index
cVAE	Conditional variational autoencoders
ESR1	Estrogen receptor 1
FFPE	Formalin-fixed paraffin embedded
fRNA-seq	FFPE-derived RNA-seq
GAT	Graph Attention Network
GEO	Gene expression omnibus
k-BET	K-nearest neighbor batch effect test
KRT5	Keratin 5
KL	Kullback–Leibler
KS	Kolmogorov–Smirnov
JSD	Jensen-Shannon divergence
PREFFECT	PaRaffin embedded formalin-FixEd cleaning tool
MRE	Median Relative Error
NB	Negative Binomial
scRNA-seq	Single cell RNA-seq
TMBC	The metastatic breast cancer project
ZI	Zero-inflated
ZINB	Zero-inflated negative binomial

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12967-025-07031-y.

Supplementary Material

Acknowledgements

We thank SA Nazlica for assistance during the creation of the software.

Author Contributions

VD and MH conceptualized the project. EJM and MH developed the statistical and computational methodology, and EJM performed the computational experiments. WZ and AS contributed to the development of the software. ST and SN-M contributed data resources and analysis of data. EJM and MH wrote the manuscript. MH and ER acquired funding. VD, ER and MH supervised the team.

Funding

We acknowledge financial support from the Natural Sciences and Engineering Research Council of Canada (NSERC, RGPIN-2018-05085) to MH, the Canadian Institutes of Health Research (CIHR, #391682) to ER and MH, and Canada Foundation for Innovation (CFI, #43481) for the computing infrastructure (VD, MH).

Data availability

PREFFECT is available at https://github.com/hallettmiket/preffect. The code used for analyses in this paper is available at https://github.com/hallettmiket/preffect-paper. It is made available under a Creative Commons Zero v1.0 Universal license. The synthetic datasets utilized in this article are available in the Zenodo repository (https://zenodo.org/records/15079531). All publicly available FFPE datasets can be found in the Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/).

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no conflicts or Conflict of interest for any aspect of this manuscript.

Author details

[1]Department of Biochemistry, Western University, 1151 Richmond St., London, ON N6A 3K7, Canada
[2]Department of Computer Science, Columbia University, 116th and Broadway, New York, NY 10027, USA
[3]Department of Radiation Oncology, Sunnybrook Health Sciences Centre - University of Toronto, 2075 Bayview Ave, North York, ON M4N3M5, Canada
[4]Department of Laboratory Medicine and Pathobiology, Sunnybrook Health Sciences Centre - University of Toronto, 2075 Bayview Ave, North York, ON M4N3M5, Canada
[5]Department of Anatomic Pathology, Sunnybrook Health Sciences Centre - University of Toronto, 2075 Bayview Ave, North York, ON M4N3M5, Canada
[6]Department of Anatomy and Cell Biology, Western University, 1151 Richmond St., London, ON N6A 3K7, Canada
[7]Department of Oncology, Western University, 1151 Richmond St., London, ON N6A 3K7, Canada

Received: 31 March 2025 / Accepted: 14 August 2025
Published online: 30 September 2025

References

1. Shi SR, Shi Y, Taylor CR. Antigen retrieval immunohistochemistry: review and future prospects in research and diagnosis over two decades. J Histochem Cytochem. 2011;59(1):13–32. https://doi.org/10.1369/jhc.2010.957191.
2. Kokkat TJ, Patel MS, McGarvey D, LiVolsi VA, Baloch ZW. Archived formalin-fixed paraffin-embedded (FFPE) blocks: a valuable underexploited resource for extraction of DNA, RNA, and protein. Biopreservation and Biobanking. 2013;11(2):101–6. https://doi.org/10.1089/bio.2012.0052.
3. Wertz DC. Archived specimens: a platform for discussion. Community Genet. 1999;2(2/3):51–60.
4. Blow N. Tissue issues. Nature. 2007;448(7156):959–60. https://doi.org/10.1038/448959a.
5. Hedegaard J, Thorsen K, Lund MK, Hein AMK, Hamilton-Dutoit SJ, Vang S, et al. Next-generation sequencing of RNA and DNA isolated from paired fresh-frozen and formalin-fixed paraffin-embedded samples of human cancer and normal tissue. PLoS ONE. 2014;9(5):e98187. https://doi.org/10.1371/journal.pone.0098187.
6. Schweiger MR, Kerick M, Timmermann B, Albrecht MW, Borodina T, Parkhomchuk D, et al. Genome-wide massively parallel sequencing of formaldehyde

fixed-paraffin embedded (FFPE) tumor tissues for copy-number- and mutation-analysis. PLoS ONE. 2009;4(5):e5548. https://doi.org/10.1371/journal.pone.0005548.
7. Wood HM, Belvedere O, Conway C, Daly C, Chalkley R, Bickerdike M, et al. Using next-generation sequencing for high resolution multiplex analysis of copy number variation from nanogram quantities of DNA from formalin-fixed paraffin-embedded specimens. Nucl Acids Res. 2010;38(14):e151–e151. https://doi.org/10.1093/nar/gkq510.
8. Tuononen K, Mäki-Nevala S, Sarhadi VK, Wirtanen A, Rönty M, Salmenkivi K, et al. Comparison of targeted next-generation sequencing (NGS) and real-time PCR in the detection of EGFR, KRAS, and BRAF mutations on formalin-fixed, paraffin-embedded tumor material of non-small cell lung carcinoma-superiority of NGS. Genes Chromosom Cancer. 2013;52(5):503–11. https://doi.org/10.1002/gcc.22047.
9. Spencer DH, Sehn JK, Abel HJ, Watson MA, Pfeifer JD, Duncavage EJ. Comparison of clinical targeted next-generation sequence data from formalin-fixed and fresh-frozen tissue specimens. J Mol Diagn. 2013;15(5):623–33. https://doi.org/10.1016/j.jmoldx.2013.05.004.
10. Weng L, Wu X, Gao H, Mu B, Li X, Wang JH, et al. MicroRNA profiling of clear cell renal cell carcinoma by whole-genome small RNA deep sequencing of paired frozen and formalin-fixed, paraffin-embedded tissue specimens. J Pathol. 2010;222(1):41–51. https://doi.org/10.1002/path.2736.
11. Sinicropi D, Qu K, Collin F, Crager M, Liu ML, Pelham RJ, et al. Whole transcriptome RNA-Seq analysis of breast cancer recurrence risk using formalin-fixed paraffin-embedded tumor tissue. PLoS ONE. 2012;7(7):e40092. https://doi.org/10.1371/journal.pone.0040092.
12. Adiconis X, Borges-Rivera D, Satija R, DeLuca DS, Busby MA, Berlin AM, et al. Comparative analysis of RNA sequencing methods for degraded or low-input samples. Nat Methods. 2013;10(7):623–9. https://doi.org/10.1038/nmeth.2483.
13. Norton N, Sun Z, Asmann YW, Serie DJ, Necela BM, Bhagwate A, et al. Gene expression, single nucleotide variant and fusion transcript discovery in archival material from breast tumors. PLoS ONE. 2013;8(11):e81925. https://doi.org/10.1371/journal.pone.0081925.
14. Wang D, Rolfe PA, Foernzler D, O'Rourke D, Zhao S, Scheuenpflug J, et al. Comparison of two illumina whole transcriptome RNA sequencing library preparation methods using human cancer FFPE specimens. Technol Cancer Res Treat. 2022;21:15330338221076304. https://doi.org/10.1177/15330338221076304.
15. Choi Y, Kim A, Kim J, Lee J, Lee SY, Kim C. Optimization of RNA extraction from formalin-fixed paraffin-embedded blocks for targeted next-generation sequencing. J Breast Cancer. 2017;20(4):393. https://doi.org/10.4048/jbc.2017.20.4.393.
16. Ludyga N, Grünwald B, Azimzadeh O, Englert S, Höfler H, Tapio S, et al. Nucleic acids from long-term preserved FFPE tissues are suitable for downstream analyses. Virchows Archiv Int J Pathol. 2012;460(2):131–40. https://doi.org/10.1007/s00428-011-1184-9.
17. Li J, Fu C, Speed TP, Wang W, Symmans WF. Accurate RNA sequencing from formalin-fixed cancer tissue to represent high-quality transcriptome from frozen tissue. JCO Precis Oncol. 2018;2:1–9. https://doi.org/10.1200/PO.17.00091.
18. Daugaard I, Kjeldsen TE, Hager H, Hansen LL, Wojdacz TK. The influence of DNA degradation in formalin-fixed, paraffin-embedded (FFPE) tissue on locus-specific methylation assessment by MS-HRM. Exp Mol Pathol. 2015;99(3):632–40. https://doi.org/10.1016/j.yexmp.2015.11.007.
19. Dietrich D, Uhl B, Sailer V, Holmes EE, Jung M, Meller S, et al. Improved PCR performance using template DNA from formalin-fixed and paraffin-embedded tissues by overcoming PCR inhibition. PLoS ONE. 2013;8(10):e77771. https://doi.org/10.1371/journal.pone.0077771.
20. Gilbert MTP, Haselkorn T, Bunce M, Sanchez JJ, Lucas SB, Jewell LD, et al. The isolation of nucleic acids from fixed, paraffin-embedded tissues-which methods are useful when? PLoS ONE. 2007;2(6):e537. https://doi.org/10.1371/journal.pone.0000537.
21. Guyard A, Boyez A, Pujals A, Robe C, Tran Van Nhieu J, Allory Y, et al. DNA degrades during storage in formalin-fixed and paraffin-embedded tissue blocks. Virchows Arch. 2017;471(4):491–500. https://doi.org/10.1007/s00428-017-2213-0.
22. Ademà V, Torres E, Solé F, Serrano S, Bellosillo B. Paraffin Treasures: Do They Last Forever? Biopreservation and Biobanking. 2014;12(4):281–3. https://doi.org/10.1089/bio.2014.0010.
23. Qq Y, Yang R, Jf S, Ny Z, Dy L, Sha S, et al. Effect of preservation time of formalin-fixed paraffin-embedded tissues on extractable DNA and RNA quantity. J Int Med Res. 2020;48(6):030006052093125. https://doi.org/10.1177/0300060520931259.
24. Do H, Dobrovic A. Sequence artifacts in DNA from formalin-fixed tissues: causes and strategies for minimization. Clin Chem. 2015;61(1):64–71. https://doi.org/10.1373/clinchem.2014.223040.
25. Ofner R, Ritter C, Ugurel S, Cerroni L, Stiller M, Bogenrieder T, et al. Non-reproducible sequence artifacts in FFPE tissue: an experience report. J Cancer Res Clin Oncol. 2017;143(7):1199–207. https://doi.org/10.1007/s00432-017-2399-1.
26. Williams C, Pontén F, Moberg C, Söderkvist P, Uhlén M, Pontén J, et al. A high frequency of sequence alterations is due to formalin fixation of archival specimens. Am J Pathol. 1999;155(5):1467–71. https://doi.org/10.1016/S0002-9440(10)65461-2.
27. Matsubara T, Soh J, Morita M, Uwabo T, Tomida S, Fujiwara T, et al. DV200 index for assessing RNA integrity in next-generation sequencing. Biomed Res Int. 2020;2020:9349132. https://doi.org/10.1155/2020/9349132.
28. Jacobsen SB, Tfelt-Hansen J, Smerup MH, Andersen JD, Morling N. Comparison of whole transcriptome sequencing of fresh, frozen, and formalin-fixed, paraffin-embedded cardiac tissue. PLoS ONE. 2023;18(3):e0283159. https://doi.org/10.1371/journal.pone.0283159.
29. Gallegos Ruiz MI, Floor K, Rijmen F, Grünberg K, Rodriguez JA, Giaccone G. EGFR and K-ras mutation analysis in non-small cell lung cancer: comparison of paraffin embedded versus frozen specimens. Anal Cell Pathol. 2007;29(3):257–64. https://doi.org/10.1155/2007/568205.
30. Vahrenkamp JM, Szczotka K, Dodson MK, Jarboe EA, Soisson AP, Gertz J. FFPEcap-seq: a method for sequencing capped RNAs in formalin-fixed paraffin-embedded samples. Genome Res. 2019;29(11):1826–35. https://doi.org/10.1101/gr.249656.119.
31. Cazzato G, Caporusso C, Arezzo F, Cimmino A, Colagrande A, Loizzi V, et al. Formalin-Fixed and Paraffin-Embedded Samples for Next Generation Sequencing: Problems and Solutions. Genes. 2021;12(10):1472. https://doi.org/10.3390/genes12101472.
32. Yin S, Wang X, Jia G, Xie Y. MIXnorm: normalizing RNA-seq data from formalin-fixed paraffin-embedded samples. Bioinformatics. 2020;36(11):3401–8. https://doi.org/10.1093/bioinformatics/btaa153.
33. Yin S, Zhan X, Yao B, Xiao G, Wang X, Xie Y. SMIXnorm: Fast and Accurate RNA-Seq Data Normalization for Formalin-Fixed Paraffin-Embedded Samples. Front Genet. 2021;12:650795. https://doi.org/10.3389/fgene.2021.650795.
34. Eraslan G, Simon LM, Mircea M, Mueller NS, Theis FJ. Single-cell RNA-seq denoising using a deep count autoencoder. Nat Commun. 2019;10(1):390. https://doi.org/10.1038/s41467-018-07931-2.
35. Grønbech CH, Vording MF, Timshel PN, Sønderby CK, Pers TH, Winther O. scVAE: variational auto-encoders for single-cell gene expression data. Bioinformatics. 2020;36(16):4415–22. https://doi.org/10.1093/bioinformatics/btaa293.
36. Yu Z, Lu Y, Wang Y, Tang F, Wong Kc, Li X. ZINB-Based Graph Embedding Autoencoder for Single-Cell RNA-Seq Interpretations. In: AAAI conference on artificial intelligence; 2022.
37. Pierson E, Yau C. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome Biol. 2015;16:241. https://doi.org/10.1186/s13059-015-0805-z.
38. Risso D, Perraudeau F, Gribkova S, Dudoit S, Vert JP. A general and flexible method for signal extraction from single-cell RNA-seq data. Nat Commun. 2018;9(1):284. https://doi.org/10.1038/s41467-017-02554-5.
39. Tian T, Min MR, Wei Z. Model-based autoencoders for imputing discrete single-cell RNA-seq data. Methods. 2021;192:112–9. https://doi.org/10.1016/j.ymeth.2020.09.010.
40. Prabhakaran S, Azizi E, Carr A, Pe'er D. Dirichlet process mixture model for correcting technical variation in single-cell gene expression data. JMLR Workshop Conf Proc. 2016;48:1070–9.
41. Svensson V. Droplet scRNA-seq is not zero-inflated. Nat Biotechnol. 2020;38(2):147–50. https://doi.org/10.1038/s41587-019-0379-5.
42. Vallejos CA, Marioni JC, Richardson S. BASiCS: Bayesian analysis of single-cell sequencing data. PLoS Comput Biol. 2015;11(6):e1004333. https://doi.org/10.1371/journal.pcbi.1004333.
43. Van Dijk D, Sharma R, Nainys J, Yim K, Kathail P, Carr AJ, et al. Recovering gene interactions from single-cell data using data diffusion. Cell. 2018;174(3):716-729.e27. https://doi.org/10.1016/j.cell.2018.05.061.
44. Huang M, Wang J, Torre E, Dueck H, Shaffer S, Bonasio R, et al. SAVER: gene expression recovery for single-cell RNA sequencing. Nat Methods. 2018;15(7):539–42. https://doi.org/10.1038/s41592-018-0033-z.

45. Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. Deep generative modeling for single-cell transcriptomics. Nat Meth. 2018;15(12):1053–8. https://doi.org/10.1038/s41592-018-0229-2.
46. Rao J, Zhou X, Lu Y, Zhao H, Yang Y. Imputing single-cell RNA-seq data by combining graph convolution and autoencoder neural networks. Science. 2021;24(5):102393. https://doi.org/10.1016/j.isci.2021.102393.
47. Veličković P, Cucurull G, Casanova A, Romero A, Liò P, Bengio Y. Graph Attention Networks 2018 Feb; https://doi.org/10.48550/arXiv.1710.10903.
48. Baul S, Ahmed KT, Filipek J, Zhang W. omicsGAT: graph attention network for cancer subtype analyses. Int J Mol Sci. 2022;23(18):10220. https://doi.org/10.3390/ijms231810220.
49. Xu C, Cai L, Gao J. An efficient scRNA-seq dropout imputation method using graph attention network. BMC Bioinform. 2021;22(1):582. https://doi.org/10.1186/s12859-021-04493-x.
50. Gayoso A, Steier Z, Lopez R, Regier J, Nazor KL, Streets A, et al. Joint probabilistic modeling of single-cell multi-omic data with totalVI. Nat Methods. 2021;18(3):272–82. https://doi.org/10.1038/s41592-020-01050-x.
51. Virshup I, Bredikhin D, Heumos L, Palla G, Sturm G, Gayoso A, et al. The scverse project provides a computational ecosystem for single-cell omics data analysis. Nat Biotechnol. 2023;41(5):604–6. https://doi.org/10.1038/s41587-023-01733-8.
52. Brody S, Alon U, Yahav E. How Attentive are Graph Attention Networks? 2022 Jan;https://doi.org/10.48550/arXiv.2105.14491.
53. Python Reference Library. Python Software Foundation. Available from: https://www.python.org.
54. R Core Team. R: A language and environment for statistical computing, vol 4.0.2. Vienna, Austria: R Foundation for Statistical Computing; 2021. Available from: https://www.R-project.org/.
55. McInnes L, Healy J, Melville J. UMAP: uniform manifold approximation and projection for dimension reduction. arXiv:1802.03426 [stat].
56. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 2018;19(1):15. https://doi.org/10.1186/s13059-017-1382-0.
57. Jain E, Zañudo JGT, McGillicuddy M, Abravanel DL, Thomas BS, Kim D, et al. The Metastatic Breast Cancer Project: leveraging patient-partnered research to expand the clinical and genomic landscape of metastatic breast cancer and accelerate discoveries. Oncology. 2023. https://doi.org/10.1101/2023.06.07.23291117.
58. de Bruijn I, Kundra R, Mastrogiacomo B, Tran TN, Sikina L, Mazor T, et al. Analysis and visualization of longitudinal genomic and clinical data from the AACR project GENIE biopharma collaborative in cBioPortal. Can Res. 2023;83(23):3861–7. https://doi.org/10.1158/0008-5472.CAN-23-0816.
59. Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol. 2009;27(8):1160–7. https://doi.org/10.1200/JCO.2008.18.1370.
60. Seabold S, Perktold J. statsmodels: econometric and statistical modeling with python. In: 9th Python in Science Conference; 2010.
61. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020;17(3):261–72. https://doi.org/10.1038/s41592-019-0686-2.
62. Doersch C. Tutorial on Variational Autoencoders 2021. arXiv:1606.05908 [stat]. https://doi.org/10.48550/arXiv.1606.05908.
63. Sethuraman J. A constructive definition of Dirichlet priors. Stat Sin. 1994;4:639–50.
64. Clarke ZA, Bader GD. *MALAT1* expression indicates cell quality in single-cell RNA sequencing data. bioRxiv 603469. 2024. https://doi.org/10.1101/2024.07.14.603469.
65. Bergholtz H, Lien TG, Swanson DM, Frigessi A, Daidone MG, Tost J, et al. Contrasting DCIS and invasive breast cancer by subtype suggests basal-like DCIS as distinct lesions. NPJ Breast Cancer. 2020;6:26. https://doi.org/10.1038/s41523-020-0167-x.
66. Zhang Y, Parmigiani G, Johnson WE. ComBat-seq: batch effect adjustment for RNA-seq count data. NAR Genom Bioinform. 2020. https://doi.org/10.1093/nargab/lqaa078.
67. Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci. 2001;98(19):10869–74. https://doi.org/10.1073/pnas.191367098.
68. Finak G, Bertos N, Pepin F, Sadekova S, Souleimanova M, Zhao H, et al. Stromal gene expression predicts clinical outcome in breast cancer. Nat Med. 2008;14(5):518–27. https://doi.org/10.1038/nm1764.
69. De Souto MC, Jaskowiak PA, Costa IG. Impact of missing data imputation methods on gene expression clustering and classification. BMC Bioinform. 2015;16(1):64. https://doi.org/10.1186/s12859-015-0494-3.

Publisher's Note