Gene expression

Per-sample standardization and asymmetric winsorization lead to accurate clustering of RNA-seq expression profiles

Davide Risso and Stefano Maria Pagnotta

1Department of Statistical Sciences, Università degli Studi di Padova, Padova, Italy and 2Department of Science and Technology, Università degli Studi del Sannio, Benevento, Italy

*To whom correspondence should be addressed.

Abstract

Motivation: Data transformations are an important step in the analysis of RNA-seq data. Nonetheless, the impact of transformation on the outcome of unsupervised clustering procedures is still unclear.

Results: Here, we present an Asymmetric Winsorization per-Sample Transformation (AWST), which is robust to data perturbations and removes the need for selecting the most informative genes prior to sample clustering. Our procedure leads to robust and biologically meaningful clusters both in bulk and in single-cell applications.

Availability and implementation: The AWST method is available at https://github.com/drisso/awst. The code to reproduce the analyses is available at https://github.com/drisso/awst_analysis

Contact: pagnotta@unisannio.it

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

Three major issues have caught most of the attention in the statistical analysis of gene expression data: (i) the classification or clustering of samples in molecularly defined classes (e.g. subtypes of a disease; Dudoit and Fridlyand, 2002; Dudoit et al., 2002), (ii) the identification of gene-level differences between groups of samples (i.e. differential expression; McCarthy et al., 2012) and (iii) the functional characterization of the differentially expressed genes in predefined categories (i.e. gene-set enrichment analysis; Geistlinger et al., 2021).

Most of the statistical contributions in these areas focus on either the identification of data preprocessing transformations to remove systematic biases (i.e. normalization; Dillies et al., 2013), on the choice of the best inferential model, such as the negative binomial model (Robinson and Smyth, 2007), or on robust and flexible clustering techniques (Risso et al., 2018). Less attention has been paid to data transformations useful to improve the outcome of unsupervised clustering procedures. This is the focus of the present article.

Data preprocessing is often the first step before the application of downstream statistical methods. To the end of meeting as close as possible the assumptions of a statistical model, data transformations may be used. Prior to the unsupervised clustering of RNA-seq data, a logarithmic transformation and a standardization step are often performed. Recent workflows proposed in the literature (e.g. Radovich et al., 2018; TCGA Research Network, 2015) differ for the order in which the two steps are performed and for the estimation of the location and scale parameters for the standardization (see Supplementary Information). One of the expected effects of such procedures is to mitigate the influence of the outlying gene expressions, but, at the same time, the logarithm can increase the variance of the many lowly expressed genes. A selection of ‘informative genes’ can ameliorate this issue empirically, but misuse of the log transformation can lead to erroneous results (Feng et al., 2014) and biases (Lun, 2018).

Many clustering procedures involve the computation of similarity/dissimilarity measures across samples and between clusters. Two recent papers (Jaskowiak et al., 2018; Vidman et al., 2019) evaluate the performance of clustering applied to complex RNA-seq datasets. They recommend (a) the average linkage for hierarchical clustering applied to correlation-based distances and (b) the Partitioning Around Medoids (PAM) (Kaufman and Rousseeuw, 1990) procedure paired with the Euclidean distance. On the other hand, applied studies (Radovich et al., 2018; TCGA Research Network, 2015) use the resampling methodology of Monti (2003). In this context, too, average linkage and correlation distance, or PAM and Euclidean distance, are common choices.

On the other hand, single-cell data are often clustered using partitioning approaches, either through the use of k-means (Kiselev et al., 2017; Risso et al., 2018), or via graph-based methods, such as the Louvain algorithm (Blondel et al., 2008), as done for instance in
the popular Seurat R package (Satija et al., 2015), Tian et al. (2019) benchmarked several such algorithms.

A critical point of any clustering pipeline is feature selection. This general procedure often suffers from a subjective choice of the genes to include in downstream analysis. Specifically, in unsupervised clustering, the implicit assumption is that only the most highly variable genes (HVG; typically 1000–2000) carry information about the samples mutual similarity. Furthermore, the choice of the variability index is also subjective.

Overall, there is a lack of clear guidelines on the optimal transformations and feature selection procedures for the clustering of RNA-seq data. In this landscape, our contribution is twofold: (a) a two-step statistical transformation: (1) standardize the original data transformation that accounts for the count nature of the data and (2) smooth the standardized values with an asymmetric function that mimic winsorization in mitigating the effect of outliers. Our transformation applies independently to each sample; in this view, it is a per-sample transformation and we name it Asymmetric Winsorization per-Sample Transformation (AWST). As for feature selection, we propose the use of Shannon’s entropy to remove uninformative genes.

2 Materials and methods

AWST aims to regularize the original read counts to reduce the effect of noise on the clustering of samples. In fact, gene expression data are characterized by high levels of noise in both lowly expressed features, which suffer from background effects and low signal-to-noise ratio, and highly expressed features, which may be the result of amplification bias and other experimental artifacts. These effects are of utmost importance in highly degraded or low input experimental samples, such as tumor samples and single cells.

AWST comprises two main steps. In the first one, namely the standardization step, we standardize the counts by centering and scaling them, exploiting the log-normal probability distribution. We refer to the standardized counts as \( z \)-counts. The second step, name the smoothing step, leverages a highly skewed transformation that decreases the noise while preserving the influence of genes to separate molecular subtypes. We define an asymmetric smoother, which is similar in spirit to winsorization (Tukey, 1962). The need for a differential treatment of expression values at the two tails of the distribution arises from the nature of the data: in fact, while the left tail is bounded by zero, the right tail depends on the amount of gene expression in the sample, the number of sequenced reads and the copy-number alterations, especially prevalent in cancer samples (Shao et al., 2019). We refer to the output of the two steps as the smoothed-counts. A further filtering method is suggested to remove those features that only contribute noise to the clustering.

2.1 Standardization step

Let us denote with \( x_{ij} \) the read counts for gene \( j \) in sample \( i \). Several authors have noted that the right tail of \( \log x_{ij} \) is akin to a Gaussian density (Gierlinski et al., 2013; Hebenstreit et al., 2011; Hoyle et al., 2002; Okrah and Corrada Bravo, 2015) (Fig. 1A). We therefore decide to anchor the standardization of the counts to this tail, by assuming that, for each sample \( i \), the right tail is well approximated by a log-normal distribution (the dashed red curve in Fig. 1A), i.e. all \( x_{ij} \) greater than the mode are assumed to come from a log-normal distribution with parameters \( \mu_i \) and \( \sigma_i, X_i = e^{\mu_i + \sigma_i Z} \), where \( Z \) is a standard Gaussian random variable. Note that we fit the distribution across genes, independently for each sample \( i \). Since our transformation is sample specific, we will omit the index \( i \) in the following, as it applies to any sample.

For each sample, we first estimate \( \mu \) with the right mode of the log-counts empirical distribution, denoted as \( \hat{\mu} \). Specifically, the right mode is estimated as the maximum of the kernel-density estimate of the empirical distribution. This implies that we estimate the location \( \hat{\mu} \) of the counts as \( e^{\hat{\mu}} \). The use of the right mode to estimate \( \mu \) allows us to minimize the influence of noise from genes with low counts in the estimation of the location \( \nu \). Then, we mirror the values on the right of the mode (the dotted curve in Fig. 1A) and we estimate the variance \( \hat{\sigma}^2 \) with the maximum likelihood estimator on the right tail and its mirrored values. To get the variance \( \hat{\sigma}^2 \) of the counts, we exploit the log-normal distribution, i.e. \( \hat{\sigma}^2 = e^{\hat{\mu}^2}(e^{\hat{\sigma}^2} - 1) \). The standardized counts, the \( z \)-counts \( z_i \), are thus defined as

\[
z_i = \frac{x_i - \hat{\mu}}{\hat{\sigma}}.
\]

A similar approach was adopted by Hart et al. (2013) to standardize the log-FPKM (Fragments Per Kilobase of transcript per Million mapped reads) values with the purpose of identifying a threshold that separates active from background genes. Both FPKM and the log transformation may lead to biases due to the influence of highly expressed genes (Bullard et al., 2010) and zero counts (Lun, 2018), respectively. Hence, we do not act on the log-FPKM transformed data. Rather, our standardization proposal is applied in the original (read count) scale. We shall note that our aim is different: while Hart uses the transformation to classify between active and nonactive genes, we use the transformation as a preliminary step for sample clustering.

2.2 Smoothing step

The second step of our proposal is a smoothing transformation applied to the \( z \)-counts. This transformation is crucial to regularize the expression of the genes affected by systematic artifacts that may bias the characterization of molecular subtypes. In fact, even when similar samples share the same subset of up-regulated genes, the level of up-regulation can change actively from sample to sample so that a compact cluster does not come to light. On the other hand, genes with a weak expression level contribute to the bias by acting as a source of noise for distance functions in high dimensions.

Briefly, we propose a smooth and skewed version of winsorization (Tukey, 1962) applied to the \( z \)-counts. To this end, we employ a smoothing function, \( T(\cdot) \), based on a modified Gaussian distribution function, made asymmetric through the variance parameter. Let \( \Phi(z) \) be the distribution function of a standard Gaussian random variable, and consider the function

\[
s(z; \sigma_0, \lambda) = \sigma_0 \cdot \left[ 1 + 2 \cdot \lambda \cdot \left( \Phi(z) - 0.5 \right) \cdot I(z > 0) \right],
\]

where \( \sigma_0 \) is a constant and \( I(\cdot) \) is the indicator function. \( s(z; \sigma_0, \lambda) = \sigma_0 \) for all the \( z \leq 0 \), and it is an increasing function for \( z > 0 \); the parameter \( \lambda \geq 0 \) controls the growth rate of the right tail. The asymmetrization mechanism follows the one adopted by Azzalini (1985) to define the skew-normal distribution. Differently from his proposal, we applied the asymmetrization device to the variance rather than to the body of the distribution.

The smoothing function is

\[
T(x; \sigma_0, \lambda) = 2 \frac{\Phi(z; \sigma(z; \sigma_0, \lambda))}{c} x \in \mathbb{R}
\]

where

\[
c = \int_{-\infty}^{\infty} \phi(z; \sigma(z; \sigma_0, \lambda))\,dz
\]

is a standardizing factor and \( \phi(\cdot) \) the probability density function of the standard Gaussian.

In Figure 1B, we plot the \( T(\cdot) \) transformation for different values of \( \lambda \). The black curve is the Gaussian probability distribution function with \( \mu = 0 \) and \( \sigma_0 = 0.075 \). The value of \( \sigma_0 \) is empirically computed so that zeroes and counts close to zero are pushed to the minimum. As \( \lambda \) increases, from 7 (blue curve) to 13 (red curve), we observe a delay (as \( \lambda \) increases) in transforming the values to the minimum. When \( \lambda = 0 \), the smoother is very close to acting as a device that dichotomizes the original values, where the mode is the
changing point. Figure 1c, in which we draw the density estimates of the smoothed values of one The Cancer Genome Atlas (TCGA) sample, shows that the majority of the values have been shrunk around -2, while the others values gradually increase up to around 4. The effect of reducing the contribution of lowly expressed genes, and of the winsorization for the highly expressed ones, is made explicit by comparing the original log counts versus the smoothed z-counts in Figure 1d. This latter graph shows that nonexpressed and lowly expressed genes are uniformly transformed to the minimum, while the highly expressed genes are pushed to the maximum. The genes with an intensity around the mode (mapped to zero) can contribute with different levels to a distance function.

2.3 Gene filtering
The effect of the smoothing function is to push outlying genes to a maximum value and genes with a low expression level to a minimum. Genes reaching the maximum across all samples are irrelevant, as well as those always assigned to the minimum value. Moreover, the features with about the same expression level can only add noise to the distance between samples without contributing meaningfully to the similarity measure. To remove the irrelevant features, and those only contributing noise, we propose a filtering procedure based on a heterogeneity index, which is possible only because AWST forces the expression level of any sample in the same range.

Since the expression levels for every sample are bound by the same maximum and minimum values, we define a partition of $K$ equally spaced intervals. The partition can be used to categorize the smoothed-counts in each sample, and then to compute the heterogeneity index $h_j$ for each of the genes. Given a cutoff value $c$, we remove all those features fulfilling the constraint $h_j < c$. As a heterogeneity index, we use the normalized Shannon's entropy

$$h_j = -\frac{1}{\log K} \sum_{k=1}^{K} p_{jk} \log p_{jk},$$

$p_{jk} = n_{jk}/n$, where $n_{jk}$ is the number of times, across all samples, that the expression level of gene $j$ falls in the $k$-th interval.

Our filtering procedure maps the samples in a less parsimonious subspace than that induced by the rule of thumb of retaining the top 1000 features (or few more), but it often yields a finer partition of the samples, as we show in Section 3.

2.4 Parameters choice
AWST has two tuning parameters: $\lambda$ and $\sigma_0$, which control the amount of smoothing. Larger $\lambda$ values push fewer points to the maximum value (Fig. 1). Similarly, larger $\sigma_0$ values lead to less shrinkage. Empirically, we found that values of $\lambda$ ranging from 5 to 13 are good choices in practice. The default value of $\sigma_0 = 0.075$ is chosen to shrink to the minimal value the low counts, more likely to be noise, while simultaneously preserving a dynamic range of moderate and high values. The two parameters both contribute to the amount of shrinkage. In all the analyses of this article, we set the two parameters to their default values ($\sigma_0 = 0.075$, $\lambda = 13$).

Similarly, our filtering step has two tuning parameters: the threshold $c$ and the number of bins $K$. The threshold $c$ controls the amount of genes that get filtered out according to the normalized entropy. An entropy of 0 indicates the genes that will not contribute any information in the classification, falling in the same bin across all samples. In practice, a small positive value (we adopt 0.1 by default) may be preferable to 0, to additionally remove those genes that contribute very little to the classification procedure.

Also the number of bins $K$ contributes to the number of features retained for downstream analyses. We empirically found 20 as a good value, being a compromise between the retention of informative features and the removal of the noninformative ones. As $K$ decreases, the intervals become larger and a higher number of genes is removed; on the contrary, when $K$ increases the filtering
Supplementary Fig. S1e). However, such procedure is inherently
2500 features after FPKM normalization works better (Fig. 2c) and TCGA (0.99; Fig. 2a); a slightly worse result is obtained with Hart pre-
processing (G = 0.82; Fig. 2b). The preprocessings of Radovich (G = 0.34; Fig. 2c) and TCGA (G = 0.07; Fig. 2d) clearly do not re-
cover the true partition. A simpler procedure that selects the top
2500 features after FPKM normalization works better (G = 0.70; Supplementary Fig. S1e). However, such procedure is inherently
sensitive to the number of selected genes. In fact, it fails when we
double the amount of features (G = 0.07; Supplementary Fig. S1f).
We observe that the performance of FPKM decreases as we increase
the number of features (Supplementary Fig. S5). Considering all fe-
tures, VST is able to recover the theoretical clustering yielding G = 1
(Supplementary Fig. S1h), as AWST. Instead, rlog retrieves a perfect
partition (G = 1) with the top 2500 features, but its performance
decreases (G = 0.67; Supplementary Fig. S1g) when all features are
included. This, coupled with the higher computational complexity
(Supplementary Table S2) suggests that VST is preferable to rlog.

The null residuals approach proposed in Townes et al. (2019) also
shows good results, with deviance residuals (G = 0.98; Supplementary Fig. S1i) outperforming Pearson residuals (G = 0.86; Supplementary Fig. S1f). See Supplementary Table S1 for a summary of the results.

Furthermore, we compare the six preprocessings with a state-of-
the-art clustering strategy for complex RNA-seq experiments, i.e.
the resampling methodology of Monti (2003), implemented in the
ConsensusClusterPlus (Wilkerson and Hayes, 2010) R package.
Briefly, these results show that only AWST and VST lead to an optimi-
mal sample partition on all the tested clustering algorithms
(Supplementary Figs 3 and 4; see Supplementary Information for more
details).

To rank the preprocessing methods independently of the cluster-
ing algorithm used, we employ a strategy based on the average sil-
houette width: by using the true cluster labels in the silhouette
calculation, we are in principle able to rank each normalization by
its ability to allow similar samples to be close to each other (more
than to samples from other groups) in the space of the normalized
gene expression. This analysis confirms that AWST, VST and rlog
(using the top 2, 500 genes) are the best performing methods, while
TCGA and Radovich perform poorly (Fig. 3 and Supplementary
Table S1).

3.2 scMixology data
A second experiment evaluates the performance of AWST in single-
cell data. We compare AWST with three transformations: log-
normalization, VST and Townes.

To distinguish the effect of a particular clustering algorithm
from that of the data transformation, we apply several clustering
algorithms: RaceID (Gru¨n, 2020), RCA (Li et al., 2017), SC3
(Kiselev et al., 2017), RSEC (Risso et al., 2018), Seurat (Sanja et al.,
2018) with resolution equal to 0.6 (low resolution) and 1.6 (high
resolution), as previously done (Tian et al., 2019). We evaluate the
combined performance of transformation and clustering on three
datasets from the scMixology experiments (Tian et al., 2019): sc_10x
(902 cells, 16 648 features and three groups), sc_10x_5cl
(3,918 cells, 11 786 features and five groups) and RNAmix_celseq2
(340 cells, 14 804 features and seven groups). See Tian et al. (2019)
for more information on the datasets. No features selection is
applied before the clustering procedures to appreciate only the effect
of the data transformation.

Figure 4 shows the distribution of the G score for each prepro-
cessing method across all clustering algorithms and datasets. VST
and AWST show the highest median score; furthermore, AWST
shows a lower variability, resulting in a greater lower quartile and
minimum score. Log-normalization and Townes’ null residuals ap-
proach lead to less accurate clustering. See Supplementary Table S3
for the values of ECA, EPC, ARI and G on all 75 experiments (data-
set, preprocessing, clustering combinations). When we apply hier-
archical clustering with Ward’s linkage to the Euclidean distance
matrix of the scMixology datasets, we obtain the highest G values,
indicating that AWST followed by hierarchical clustering with Ward
linkage and CH-curve is an effective clustering method for single-
cell data.

3.3 Lower Grade glioma (LGG) study
A collection of 516 primary tumor samples of the LGG study from
TCGA allows to demonstrate the advantages of AWST on a com-
plex RNA-seq experiment. We compare AWST results with those
previously published, in which ad-hoc and sometimes multimomic approaches were used. While there is no ground truth in this dataset, comparing our partition to those identified by several independent, multimomic approaches, allows us to check if the derived clusters represent biologically meaningful subtypes.

LGG is classified in three molecular subtypes connected to DNA lesions (‘Original subtype’ in the plots). The first difference concerns the Isocitrate DeHydrogenase genes 1 and 2 (IDH) status. The wild-type of both genes defines an autonomous molecular subtype (IDHwt); in case of mutations, the codeletion of chromosomes 1p/19q discriminates between CODEL (IDHmut-codel) and NONCODEL (IDHmut-noncodel) molecular subtypes (Eckel-Passow et al., 2015). Ceccarelli et al. (2016) showed that the IDHmut-noncodel molecular subtype is equivalent to the G-Cimp subtype of Noushmehr et al. (2010).

We apply AWST to the RSEM estimated counts after GC-content correction and full-quantile normalization. The gene-filtering step retains 10,212 genes out of the original 19,138. We compute the Euclidean distance followed by hierarchical clustering with Ward’s linkage (Fig. 5). The lack of meaningful clusters in the dendrogram of the 8926 filtered out genes (Supplementary Fig. S11) indicates that the removed features only contribute noise.

The dendrogram in Figure 5 is annotated with different partitions. The expression clusters denoted by the ‘exprClust/CELL’ bar come from a pan-glioma study, in which stringent filtering was applied to select 2275 features in a set of 667 samples (154 GBM and 513 LGG) (Ceccarelli et al., 2016), leading to four distinct groups. A second expression clustering (‘exprClust/NEJM’) is from TCGA Research Network (2015) and provides a different partition in four groups. Kamoun et al. (2016) studied 156 oligodendrogial tumors of the IDHmut-CODEL subtype. With a multimomic approach, they found three molecular subtypes (indicated by the ‘CODEL/subtype’ bar in Fig. 5).

The CH-curve suggests a primary partition (‘clust5’) of five groups that mostly mirrors the Original subtype of brain cancer: c51 is the cluster of the IDHwt samples; c53 includes the IDHmut-codel samples. The IDHmut-noncodels splits into two groups: c54 and c55. Cluster c52 mostly corresponds to the LGr2 (Ceccarelli et al., 2016) and R4 groups (TCGA Research Network, 2015); the genomic determinants of this group are still unexplored. Clust5 is supported by the study of the overall survival time associated with each sample (Supplementary Fig. S6; log-rank test P-value of ≤2.2 × 10^{-18}).

C53 is a large collection of IDHmut-codel samples. Kamoun et al. (2016) specifically studied this kind of tumors and obtained three subgroups, supported by different genomic analyses. Our hierarchical clustering also splits c53 into three subgroups that match the corresponding ones of Kamoun. To support the identification of our subgroups, we estimated the survival curves associated with them (Supplementary Fig. S8; log-rank test P-value of 0.02). Our
finding is comparable to Supplementary Figure S8a of Kamoun et al. (2016). C54 is an almost homogeneous cluster of G-Cimp-high samples. Our clustering identifies c541 and c542. The latter matches the G-Cimp-low subtype discovered in Ceccarelli et al. (2016), and obtained using a supervised analysis of methylation data. Our experiment identifies the G-Cimp-low solely based on the expression profiles, for the first time to the best of our knowledge (Supplementary Fig. S9; log-rank test $P$-value $= 2 \times 10^{-4}$).

C55 hides a promising subtype candidate. Our hierarchical clustering splits this group into c551 and c552. While c551 contains the majority of c55 samples, c552 matches a suggestive clinical trait.
Supplementary Figure S10 indicates that c552 is associated with the best survival curve. This subgroup appears as well defined as c542, the G-CIMP-low. Although this cluster contains only few samples, the very good survival of this group makes it a good candidate for further molecular characterization.

To compare the performance of AWST to its competitors, we apply the other methods to the LGG dataset. None of the 516 LGG samples’ estimated partitions consistently match the subtypes known from the literature (Supplementary Figs S12, S14 and S16). Reassuringly, we note that our implementation of the TCGA protocol reproduces the original results of TCGA Research Network (2015), denoted in the figures by ‘exprClust/NEJM’.

Special attention has to be devoted to VST that appears the direct competitor of AWST from the synthetic analyses. Supplementary Figures 17 and 18 show the dendrograms obtained with VST-normalized data. In Supplementary Figure S17, we use the most HVG in the same number suggested by AWST; in Supplementary Figure S18, we use the same features used for Figure 5; hence, this result depends on the feature selection procedure of AWST. By qualitatively comparing the dendrogram of Figure 5 with that of Supplementary Figure S17, we observe that the molecular hierarchy of LGG is not recovered by VST: the rightmost branch collects the majority of the CODELs with a subgroup of NONCODELs. In contrast, AWST groups the CODELs samples in the same branch, matching the effect of the DNA lesion. The G-CIMP-low are well segregated (Supplementary Fig. S17), but the c552 do not explicitly form a group when using VST.

When using the AWST feature selection procedure prior to VST (Supplementary Fig. S18), the clustering better matches the molecular characterization of LGG. Furthermore, both c542 and c552 are well separated (Supplementary Fig. S18). Taken together, these analyses indicate that AWST yields a partition that matches the molecular events at the DNA level and suggest new candidate subtypes even in data already extensively analyzed. AWST can also support analyses with VST providing a good feature selection method.

The null residuals approach of Townes et al. (2019) does not perform well in the LGG data (Supplementary Fig. S19). When using the features selected by AWST, the clustering is weakly comparable with that in Figure 5 (Supplementary Fig. S20), confirming the importance of feature selection. However, c542 and c552, two small but important clusters identified by AWST, cannot be clearly found and are mixed with samples of different nature.

3.4 Cord blood mononuclear cells (CBMC)

To further demonstrate the performance of AWST, we apply the transformation to the CBMC single-cell experiment of Stockius et al. (2017) used to introduce the CIT-Seq methodology. Briefly, CIT-Seq allows independently profiling of scRNA-seq and a panel of antibodies for each cell. The antibody-derived tags data (ADTs) measure the levels of those membrane proteins able to classify each cell to the subpopulations of the immune system, essentially providing the same informative content of flow cytometry. As an example, the combination of CD3+ and CD19− is a marker of T-cells, while CD3− and CD19+ characterizes B-cells. Here, we use ADT as an independent annotation of our estimated partition, which is solely based on scRNA-seq expression.

Before applying AWST, we retain the cells having at least 500 detected genes. This preprocessing is a typical step in single-cell analysis, in which low-quality samples may be disrupted or dying cells, or empty droplets (Lun et al., 2019). The total number of cells is reduced to 7613 from the original size of 7985 (about 5% removed). AWST is applied to a data matrix of 7613 cells measured on 17014 features after full-quantile normalization. The gene-filtering step (with default parameters) returns 230 genes. None of the 230 genes retained by the filtering step is included in the set of antibodies in the ADT data.

Applying hierarchical clustering with Ward’s linkage and Euclidean distance, we obtained a partition of six clusters (Fig. 5a), each of which is, for the most part, interpretable according to the levels of expression of the independent ADT data. For instance, cluster CBMC1 identifies the T-cells subpopulation given that CD3 is highly expressed, while CD19 is low. The opposite expression of these two markers (CD3 low and CD19 high) identifies the B-cells, clustered in CBMC2. The monocyes are the cells in CBMC5 where the markers CD14 and CD11c are both expressed. Each of the six clusters shows compact subgroups that are not explained by the ADT, suggesting that the scRNA-seq data may lead to more granularity in the clustering compared to the relatively few ADT markers.

The inferred clusters are well separated in the space of the first two principal components computed from the 230 features after AWST (Fig. 5b). A three-dimensional view of the cells is at https:// turtl.com/hrswhbd. Overall, there is good agreement between ADT data and scRNA-seq-based unsupervised clustering (Supplementary Figs S21 and S22).

Generally, t-distributed Stochastic Neighbor Embedding (t-SNE; Fig. 5c) mapping (van der Maaten and Hinton, 2008) and Uniform Manifold Approximation and Projection (UMAP; Fig. 5d) (McInnes et al., 2018) are the preferred visualization methods for single-cell data. The hierarchical clustering of Figure 5a shows its superiority to unveil the complex similarity between cells. From t-SNE or UMAP, the landscape of this collection of cells appears quite flat, while the hierarchical clustering is able to unveil the relations between cell types. Overall, this analysis demonstrates that AWST leads to biologically interpretable results with a simple hierarchical clustering procedure in single-cell data.

Next, we employ these data to explore the feature selection step of our procedure and to compare it to two popular methods presented in the single-cell literature: namely, the deviance criterion of Townes et al. (2019) and a procedure based on HVG (Lun et al., 2016). Townes’ method and AWST lead to very similar gene rankings, while Lun’s method, as implemented in the scran Bioconductor package (Lun et al., 2016), selects a different subset of genes (Supplementary Fig. S26). AWST clustering with scran’s selected features is less successful at recapitulating the partition defined by the ADTs (Supplementary Fig. S23).

Finally, we compare AWST with its two main competitors, VST and Townes. VST shows a fairly good agreement with AWST (Supplementary Fig. S24). However, it fails to correctly separate myeloid DCs from monocytes, as confirmed by the ADT markers, especially when applied on the features selected by scran (Supplementary Fig. S24c). Townes is very sensitive to the feature selection step: it fails to identify cell populations when used with HVG or genes selected by scran (Supplementary Figs S25a and S25c). When applied with scry, the dendrogram better represents the cell population structure, albeit with some inconsistencies, e.g. monocytes and myeloid DCs are grouped together (Supplementary Fig. S25d). Finally, when applied on AWST selected genes, it leads to a much more consistent dendrogram than the other cases (Supplementary Fig. S25b).

Taken together, these analyses indicate that AWST yields a good cell partition without the need for complex clustering procedures. Furthermore, our analysis highlights the importance of feature selection, showing how AWST can also be used to select a robust set of genes for the downstream cluster analysis.

4 Discussion

Data transformations are an important step prior to unsupervised clustering. However, not much emphasis is usually given to this step, with most workflows simply normalizing the data for sequencing depth and employing the log transformation. Here, we have shown how this is suboptimal, and how bespoke transformations, such as AWST, VST or null residuals (Townes et al., 2019) are preferable.

We have presented a novel data transformation, AWST, with the aim of reducing the noise and regularizing the influence of outlying genes in the clustering of RNA-seq samples. AWST is a per-sample transformation that comprises two main steps: the standardization step exploits the log-normal distribution to bring the distribution of each sample on a common location and scale; the smoothing step employs an asymmetric winsorization to
push the lowly expressed genes to a minimum value and to bound the highly expressed genes to a maximum.

Unlike VST and Townes that require the whole dataset to transform the data, AWST is a per-sample transformation, in the sense that no information from the other samples in the cohort is used when transforming the data; instead, each sample is processed independently. As a consequence, AWST can be easily applied ‘out-of-sample’, e.g. on new data in a supervised classification setting, without the need of ‘freezing’ the parameters.

Our optional filtering procedure involves a heterogeneity index, Shannon’s entropy, that does not need a preliminary location estimate, unlike the variance and its robust alternatives. This strategy is possible because of the smoothing step, which maps the values of all samples into the same interval.

Applying our transformation to synthetic and real datasets, including bulk and single-cell RNA-seq experiments, shows that AWST leads to biologically meaningful clusters. We have used synthetic and real data to compare AWST with several state-of-the-art approaches. AWST is always better than or comparable to alternative approaches in all datasets.

In an LGG real dataset, AWST followed by hierarchical clustering with Ward’s linkage and Euclidean distance uncovers cancer subtypes previously discovered in independent, sometimes multitemic, experiments.

Applied to the CBMC single-cell data, AWST leads to compact, biologically meaningful clusters, consistent with independent external data. Hence, AWST offers a unified standardization framework applicable to both bulk and single-cell datasets.

In the synthetic data (Supplementary Table S4), as well, and future work includes the adaptation of the method for the characteristic shape of microarray data (cf. background signal and signal saturation).

AWST requires the choice of a few tuning parameters. While we determine the parameter values empirically in this work, we have found that the default values that we provide in the software implementation work well in a variety of cases. In fact, we use the same values for all the datasets analyzed in the paper. Potentially, one could increase the number of bins in our feature selection procedure to retain more genes if the default value yields too few genes in single-cell applications. However, we have found that the default values work well in both bulk and single-cell applications.

Furthermore, we have performed a robustness analysis on the synthetic dataset. As discussed in Section 2.4, the most important parameter is $\lambda$, which weights the importance of the features around the mode. By decreasing $\lambda$ to 0, the overall performance decreases, but it remains high ($G = 0.90$; Supplementary Table S4). The other two main parameters, $\sigma_0$ and the number of bins in the filtering procedure, have little to no effect in the performance of the clustering in the synthetic data (Supplementary Table S4).

The AWST procedure can in principle be applied to the raw RNA-seq counts. In fact, this is how it is applied to the SEQC synthetic dataset. However, we have found that in real, more complex datasets a preliminary normalization to account for sequencing depth is beneficial. We recommend a thorough exploratory data analysis to identify the type and extent of systematic bias in the data and the use the full-quantile method (Ballard et al., 2010) for this preliminary step.

AWST is computationally efficient: the most expensive step is the kernel-density estimation of the empirical distribution. We note that if the data have been normalized with the full-quantile normalization (Ballard et al., 2010) this estimation needs to happen only once, as the normalization forces the sample distributions to have the same quantities. Even without leveraging this feature, AWST is the most efficient preprocessing method among those that work well in our comparisons (Supplementary Table S2).

Overall, AWST leads to better and more robust sample segregation allowing to move from standard statistical discovery to a more in-depth investigation of subgroups within clusters on the way to precision medicine. Furthermore, AWST allows to use statistical coherent methods, such as the Ward’s linkage method applied to Euclidean distance, (both related to the ANOVA setup), which is the only guarantee for the trust-worthiness of the results in unsupervised settings.

Acknowledgments

The authors thank Chiara Romualdi and Michele Ceccarelli for fruitful discussions and for their feedback on the manuscript. The authors thank the Associate Editor and the three referees for providing thoughtful feedback.

Funding

This work was supported by Programma per Giovani Ricerconeri Rita Levi Montalcini granted by the Italian Ministry of Education, University and Research, the Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation [CZF2019-002443] and the National Cancer Institute of the National Institutes of Health [2U24CA180996] to D.R., and the Department of Science and Technology, Università degli Studi del Sannio, Benevento 82100, Italy and the AIRC Foundation under I.D.21846 Project to S.M.P.

Conflict of Interest: the authors declare no conflict of interests.

References

Anders, S. and Huber, W. (2010) Differential expression analysis for sequence count data. Genome Biol., 11, R106.
Azzalin, A. (1985) A class of distributions which includes the normal ones. Scand. J. Stat., 12, 171–178.
Blondel, V.D. et al. (2008) Fast unfolding of communities in large networks. J. Stat. Mech., 2008, P10008.
Ballard, J.H. et al. (2010) Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments. BMC Bioinform., 11, 94.
Calinski, T. and Harabasz, J. (1974) A dendrite method for cluster analysis. Commun. Stat., 3, 1–27.
Ceccarelli, M. et al. (2016) Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. Cell, 164, 550–563.
Dillies, M.A. et al. (2013) A comprehensive evaluation of normalization methods for illumina high-throughput RNA sequencing data analysis. Brief. Bioinform., 14, 671–683.
Dudoit, S. and Fridlyand, J. (2002) A prediction-based resampling method for estimating the number of clusters in a dataset. Genome Biol., 3, research0036.1–research0036.21.
Dudoit, S. et al. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. J. Am. Stat. Assoc., 97, 77–87.
Eckel-Passow, J.E. et al. (2015) Glioma groups based on 1p/19q IDH, and TERT promoter mutations in tumors. N. Engl. J. Med., 372, 2499–2508.
Feng, C. et al. (2014) Log-transformation and its implications for data analysis. Shanghai Arch. Psychiatry, 26, 105–109.
Geistlinger, L. et al. (2021) Toward a gold standard for benchmarking gene set enrichment analysis. Brief. Bioinform., 22, 545–556.
Gierlinski, M. et al. (2015) Statistical models for RNA-seq data derived from a two-condition 48-replicate experiment. Bioinformatics, 31, 3625–3630.
Grun, D. (2020) RaceID: Identification of Cell Types and Inference of Lineage Trees from Single-Cell RNA-Seq Data. R package version 0.2.1. https://cran.r-project.org/package=RaceID.
Hart, T. et al. (2013) Finding the active genes in deep RNA-seq gene expression studies. BMC Genomics, 14, 778–778.
Helenstreit, D. et al. (2011) RNA sequencing reveals two major classes of gene expression levels in metazoan cells. Mol. Syst. Biol., 7, 497–497.
Hoyle, D.C. et al. (2002) Making sense of microarray data distributions. Bioinformatics, 18, 576–584.
Hubert, L. and Arabie, P. (1985) Comparing partitions. J. Class., 2, 193–218.
Jaskowiak, P.A. et al. (2018) Clustering of RNA-seq samples: comparison study on cancer data. Methods, 132, 42–49.
Kamoun, A. et al. (2016) Integrated multi-omics analysis of oligodenrogial tumours identifies three subgroups of 1p/19q co-deleted gliomas. Nat. Commun., 7, 11263.
Kaufman, L. and Rousseuw, P.J. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley, Hoboken, NJ.
Kislev, V.Y. et al. (2017) Sc3: consensus clustering of single-cell RNA-seq data. Nat. Methods, 14, 483–486.
Li, H. et al. (2017) Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. Nat. Genet., 49, 708–718.

Love, M. et al. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with deseq2. Genome Biol., 15, 550.

Lun, A. (2018) Overcoming systematic errors caused by log-transformation of normalized single-cell RNA sequencing data. bioRxiv, 10.1101/404962.

Lun, A. et al. (2016) A step-by-step workflow for low-level analysis of single-cell RNA-seq data with bioconductor. F1000Research, 5, 2122.

Lun, A. T. et al. (2019) Emptydrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. Genome Biol., 20, 63.

McCarthy, D. J. et al. (2012) Differential expression analysis of multifactor RNA-seq experiments with respect to biological variation. Nucleic Acids Res., 40, 4288–4297.

McInnes, L. et al. (2018) UMAP: uniform manifold approximation and projection for dimension reduction. arXiv:1802.03426. 10.1038/nbt.4314.

Monti, S. (2003) Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. Mach. Learn., 52, 91–118.

Noushmehr, H. et al. (2010) Identification of a CPG island methylator phenotype that defines a distinct subgroup of glioma. Cancer Cell, 17, 510–522.

Okrah, K. and Corrada Bravo, H. (2015) Shape analysis of high-throughput transcriptomics experiment data. Biostatistics, 16, 627–640.

Radovich, M. et al. (2018) The integrated genomic landscape of thymic epithelial tumors. Cancer Cell, 33, 244–258.e10.

Risso, D. et al. (2018) clusterexperiment and rse: a bioconductor package and framework for clustering of single-cell and other large gene expression data-sets. PLoS Comput. Biol., 14, e1006378.

Robinson, M. D. and Smyth, G. K. (2007) Moderated statistical tests for assessing differences in tag abundance. Bioinformatics, 23, 2881–2887.

Rousseeuw, P. J. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math., 20, 53–65.

Satija, R. et al. (2015) Spatial reconstruction of single-cell gene expression data. Nat. Biotechnol., 33, 495–502.

Shao, X. et al. (2019) Copy number variation is highly correlated with differential gene expression: a pan-cancer study. BMC Med. Genet., 20, 175.

Stoeckius, M. et al. (2017) Simultaneous epitope and transcriptome measurement in single cells. Nat. Methods, 14, 865–868.

Su, Z. et al. (2014) A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the sequencing quality control consortium. Nat. Biotechnol., 32, 903.

TCGA Research Network (2015) Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. N. Engl. J. Med., 372, 2481–2498.

Tian, L. et al. (2019) Benchmarking single cell RNA-sequencing analysis pipeline using mixture control experiments. Nat. Methods, 16, 479–487.

Townes, F. W. et al. (2019) Feature selection and dimension reduction for single-cell RNA-seq based on a multinomial model. Genome Biol., 20, 285.

Tukey, J. W. (1962) The future of data analysis. Ann. Math. Stat., 33, 1–67.

van der Maaten, L. and Hinton, G. (2008) Visualizing data using t-SNE. J. Mach. Learn. Res., 9, 2579–2605.

Vidman, L. et al. (2019) Cluster analysis on high dimensional RNA-seq data with applications to cancer research—an evaluation study. PLoS One, 14, e0219102.

Wilkerson, M. D. and Hayes, D. N. (2010) Consensusclusterplus: a class discovery tool with confidence assessments and item tracking. Bioinformat., 26, 1572–1573.