OXFORD

Briefings in Bioinformatics, 2022, **23(5)**, 1–14

https://doi.org/10.1093/bib/bbac158

Problem Solving Protocol

Balancing the transcriptome: leveraging sample similarity to improve measures of gene specificity

Leroy Bondhus, Roshni Varma†, Yenifer Hernandez† and Valerie A. Arboleda

Corresponding author. Valerie A. Arboleda, 615 Charles E. Young Drive South, Los Angeles, CA 90095, USA. Tel.: +1-310-983-3568;
E-mail: varboleda@mednet.ucla.edu

†Roshni Varma and Yenifer Hernandez contributed equally to this work.

Abstract

The spatial and temporal domain of a gene's expression can range from ubiquitous to highly specific. Quantifying the degree to which this expression is unique to a specific tissue or developmental timepoint can provide insight into the etiology of genetic diseases. However, quantifying specificity remains challenging as measures of specificity are sensitive to similarity between samples in the sample set. For example, in the Gene-Tissue Expression project (GTEx), brain subregions are overrepresented at 13 of 54 (24%) unique tissues sampled. In this dataset, existing specificity measures have a decreased ability to identify genes specific to the brain relative to other organs. To solve this problem, we leverage sample similarity information to weight samples such that overrepresented tissues do not have an outsized effect on specificity estimates. We test this reweighting procedure on 4 measures of specificity, Z-score, Tau, Tsi and Gini, in the GTEx data and in single cell datasets for zebrafish and mouse. For all of these measures, incorporating sample similarity information to weight samples results in greater stability of sets of genes called as specific and decreases the overall variance in the change of specificity estimates as sample sets become more unbalanced. Furthermore, the genes with the largest improvement in their specificity estimate's stability are those with functions related to the overrepresented sample types. Our results demonstrate that incorporating similarity information improves specificity estimates' stability to the choice of the sample set used to define the transcriptome, providing more robust and reproducible measures of specificity for downstream analyses.

Keywords: gene specificity, similarity, weighting, transcriptomics, GTEx

Introduction

The transcriptome is the set of potential or realized states of gene expression in a cell, tissue or organism. In human adults, there are estimated to be over 400 distinct cell-types that each develop along a unique developmental trajectory [1]. Add to this the diversity of progenitor cells and intermediate transition cell-states that occur earlier in development, and one begins to appreciate the complexity of information relayed through the transcriptome. To guide development through this diversity of cell-types and states requires the ubiquitous expression of genes with global functions for cell proliferation and survival as well as the precise expression of genes that control specialized developmental programs. The full extent of a gene's functions is not known *a priori*, so investigating spatial and developmental patterns of gene expression, i.e. the context of expression, can provide insight into the gene's function. This context of gene expression can partially explain the phenotype that results when a given gene is mutated ([2–5]) or be used to investigate whether the gene is involved in the specialized functions of a given cell, tissue or developmental event and what those specialized functions might be [6–9]. A useful summary of the degree to which a gene's expression leans toward ubiquity or specialization is the aim of gene expression specificity measurements.

While methods for quantifying gene expression are well established [10], measuring the specificity of gene expression requires addressing additional challenges (for a review of current methods of measuring specificity of gene expression see [11]). We address here one emergent challenge associated with the choice of a transcriptomic data sample set on which to measure specificity. Often, tissues and organs can be subdivided in numerous ways, such as dividing the brain into distinct functional domains or along different developmental axes, which often include gradients of expression changes rather than discrete transition points. In the brain, the transcription profiles of these subregions tend to be highly correlated with one another, reflecting the common functions and developmental origins of these subregions [12]. This leads to a problem, however, as measures of specificity are sensitive to the sampling depth into any

Leroy Bondhus is a PhD student in the Genetics and Genomics Home Area at the University of California Los Angeles.
Roshni Varma is an undergraduate student at the University of California Los Angeles.
Yenifer Hernandez is an undergraduate student at the University of California Los Angeles.
Valerie A. Arboleda is an Assistant Professor in the Departments of Computational Medicine, of Pathology and Lab Medicine and of Human Genetics.

Received: January 10, 2022. **Revised:** April 6, 2022. **Accepted:** April 10, 2022

particular organ system or timepoint [11]. A consequence of this sensitivity is that the ability of measures of gene specificity to detect genes specific to regions or timepoints is diminished if they are highly similar to other regions or timepoints that are overrepresented in the sample set. In the case of the brain, this means that sampling multiple brain subregions decreases a specificity measure's ability to detect brain-specific genes.

One potential means of alleviating the problems associated with adding sampling depth to a particular organ system when building a representative sample set is by using sample similarity information to weight samples to adjust each sample's contribution to the measure of specificity. To establish the intuition for this, consider a sample set that includes biological replicates. Biological replicates of the same sample type tend to have very high similarity, and therefore, the weight of any given replicate should be inversely proportional to the number of replicates coming from the same sample type. By extension, the weight of individual samples from different regions of a common organ should be inversely related to the number of regions sampled from that organ. Collectively, these examples point to the intuition that the weight assigned to a sample should be inversely related to its similarity to the other samples in the sample set, suggesting that sample similarity is a natural metric on which to assign sample weight for measures such as specificity.

Presently, no existing methods for measuring gene specificity take into account the similarity between samples in the sample set used to define the transcriptome. This leads to instability in existing measures of specificity when called on datasets that vary in the depth of sampling of particular biological contexts. Here, we propose a generalizable procedure for integrating sample similarity information with measures of gene specificity and demonstrate how this natural integration of sample similarity information stabilizes specificity measures.

Results

Description of problem and proposed solution

The similarity between cells and tissues can distort measures of specificity for gene expression. A balanced sample set where all sampled tissues share nearly the same level of similarity with one another facilitates specificity measures that match intuition (Figure 1A). However, balancing the sample set by considering only tissues that are at approximately the same level of similarity occurs at the expense of deeper sampling of individual tissue subregions. Adding depth to one sample type (e.g. brain), without an equivalent addition of sample types for other organs, can substantially change the measured degree of specificity of gene expression. This problem is demonstrated in the toy example in Figure 1A, where the brain marker *OLIG1* has specificity comparable to other marker genes, such as *PRSS2* in the pancreas and *MYH6* in the heart, when the sample set is relatively balanced, but when the sample set is expanded to include samples from additional brain subregions, *OLIG1* ceases to appear specific to any brain tissues.

For the present study, we compare several measures of gene specificity that are amenable to incorporation of sample weights. These measures are Z-score [13], Tau [14], Tsi [15] and Gini [16, 17] which were previously compared in a benchmarking study of measures of gene specificity [11] (see Methods for details). We chose to look at these several measures of specificity to test whether incorporation of sample similarity information to measures of gene specificity could improve a variety of different measures. Throughout the manuscript, we refer to raw specificity measures that do not incorporate sample similarity information as *flat* measures and measures that do incorporate sample similarity information via weights as *weighted* measures.

To test the effect of incorporating sample similarity derived weights on measures of specificity, we used three RNA-seq datasets including a human tissue sample set from the Gene-Tissue Expression (GTEx) project [18] and single cell datasets from zebrafish [19] and mouse [20]. From GTEx, the matrix of the median gene expression values across all individuals for each of 54 unique tissue types was used. Brain-region samples are overrepresented in the GTEX dataset, making up 24% (13/54) of the different available tissue samples. This enabled us to explore how specificity values varied when measured on balanced sample sets where only one tissue from each organ system was included compared with unbalanced sample sets where there is an overrepresentation of brain regions. The zebrafish single cell dataset comes from [19] and includes 220 unique cell clusters from four developmental time points, subsets of which are used for our analyses. The mouse single cell dataset comes from [20] and includes cells from 98 major cell clusters representing over 50 mouse tissues and cultures: subsets of these clusters were used for our analyses. As a note, we use the term *sample* throughout to refer to either unique tissues or cell clusters as opposed to biological or technical replicates.

To test the effects of measuring specificity on an unbalanced sample set, we looked first at the correlation of specificity values measured on sample sets that were either balanced or unbalanced with respect to the set of tissues or cell clusters included. In the balanced GTEx subsets, we include only one brain subregion in the sample set compared with the unbalanced GTEx subset that includes all brain-region samples in the sample set. The correlation between balanced and unbalanced sample sets was repeated for each choice of brain subregion for the balanced sample set. All nonbrain samples were included in both the balanced and unbalanced sample sets. Genes with Z-score greater than 2 were considered specific with higher values indicating greater specificity. Tau, Tsi and Gini measures are all on the scale between 0 and 1, where 0 indicates

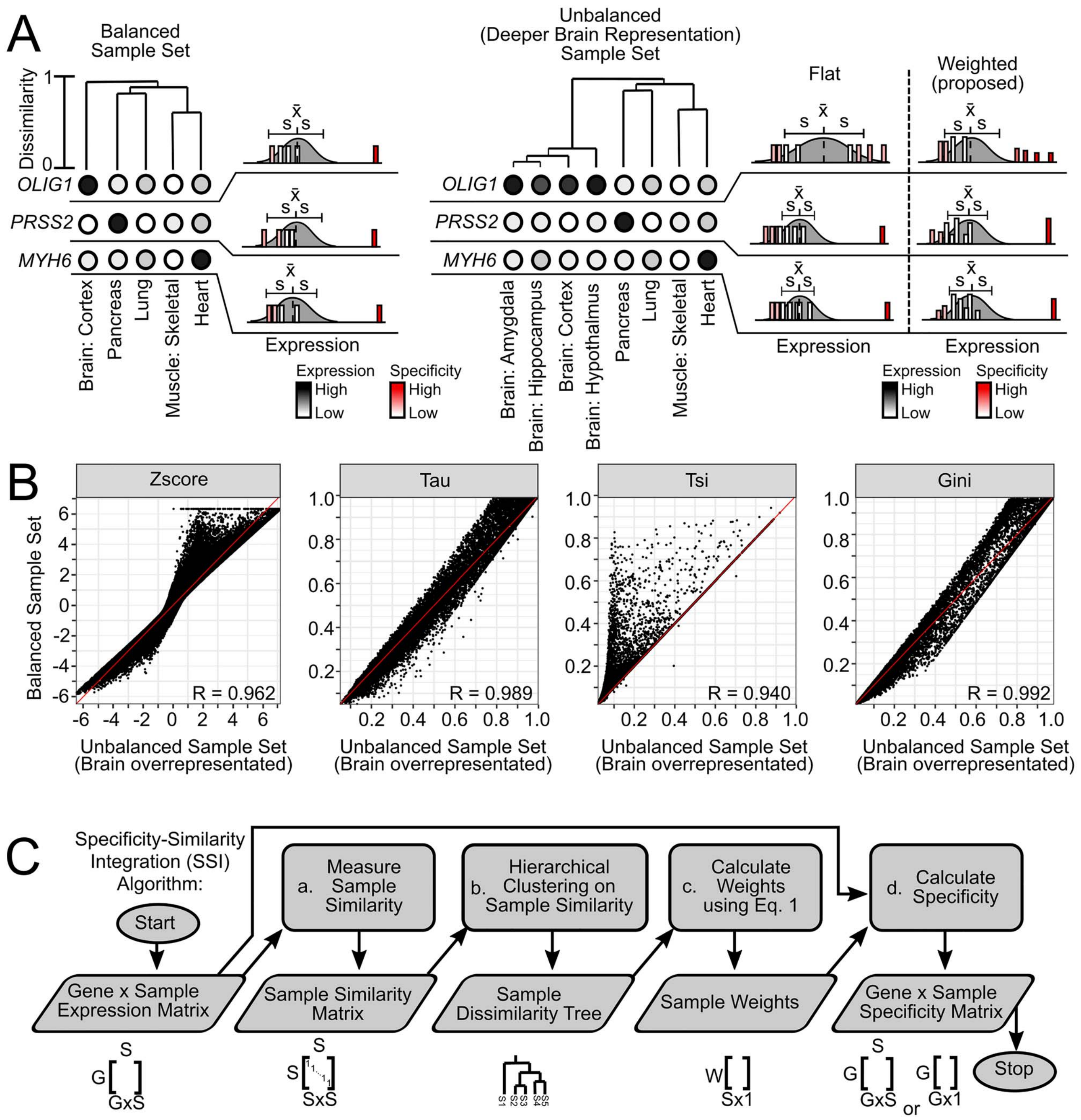

Figure 1. Problem with unbalanced sample sets for measuring gene specificity and the proposed solution. (**A**) Toy diagram of problem addressed. Global (dis)similarity of tissues is represented as a dendrogram for the balanced sample set (left) and the unbalanced sample set (right) that has an excess of brain subregions included. The color of each dot represents the relative expression of the gene in the given tissue sample. Fitted normal curve is shown to the right with sample mean (x̄) and sample SD (s) for log expression values. Bars plotted with the fitted normal curves each represent an individual tissue sample's expression and the bar's relative height represents that sample's relative weight. Specificity, as measured by the Z-score, is the number of SD of the bar from the sample mean for the given gene-sample pair and is represented by the color of the bar. (**B**) Change in specificity measures with deeper brain sampling. On x- and y-axes are specificity values measured on the unbalanced and balanced sample sets, respectively, for each gene (or gene-tissue pair for Z-score). The unbalanced sample set includes all nonbrain samples and all brain subregion samples, while the balanced sample set includes all nonbrain and one randomly selected brain subregion sample from the GTEx dataset. (**C**) SSI Algorithm or workflow proposed for integrating sample similarity information into specificity measures.

nonspecific expression and 1 indicates the maximum specificity in the tissue or cell type with the highest overall level of expression for that gene. Overall, we observe a strong correlation ($R > 0.9$) between specificity values calculated using the balanced sample set and those calculated using the unbalanced sample set for all measures of gene specificity with the highest average correlation observed for Gini $R = 0.991$ (95% CI: 0.991–0.991, one sample *t*-test) and lowest for Tsi $R = 0.939$ (95% CI: 0.938–0.940, one sample *t*-test). For Z-score, the average correlation was $R = 0.962$ (95% CI: 0.961–0.963, one sample *t*-test), and for Tau, $R = 0.989$ (95% CI: 0.989–0.989, one sample *t*-test). A representative example of the specificity scores measured on the balanced and unbalanced sample sets is shown for each measure in Figure 1B. Relatively strong correlations were similarly found in the single cell datasets.

However, the genes with the largest difference in specificity scores measured between the balanced and unbalanced sample sets are not a random subset of genes. For example, in the GTEx dataset, the top 1% of genes with the greatest positive differences in each specificity score between the balanced and unbalanced sample sets (i.e. where specificity in the balanced sample set is greater than specificity in the unbalanced sample set) are highlighted in Supplementary Figure S1 (see Supplementary Data available online at http://bib.oxfordjournals.org/) in red and represent genes that are the most variable as the sample set becomes unbalanced. Gene ontology enrichment analysis performed on these genes showed substantial enrichment in genes that function in brain-related processes (Supplementary Figure S1, see Supplementary Data available online at http://bib.oxfordjournals.org/). There was less consistency in terms associated with the genes where specificity in the balanced sample set was set less than specificity in the unbalanced sample set across measures (Supplemental Figure S2). The enrichment of brain-related terms in the top 1% of genes with the greatest positive difference between the balanced and unbalanced sample sets highlights how overrepresentation of particular sample types can introduce systematic biases into measures of specificity reducing the power of these measures to identify genes specific to the overrepresented sample type.

To address this problem, we decided to leverage sample similarity information to reweight samples in the sample set defining the transcriptome such that similar samples tend to share their weight while more distinct samples tend to retain more of their full weight. The general workflow proposed, which we call the Specificity-Similarity Integration (SSI) procedure, is given in Figure 1C and discussed in detail in Methods. In the GTEx dataset, we measured tissue–tissue similarity using each tissue's respective gene expression profile and found brain samples clustered together with a high degree of intragroup similarity, though cerebellar samples had a lesser degree of similarity than other brain regions (Supplementary Figure S3, see Supplementary Data available online at http://bib.oxfordjournals.org/). Following the SSI procedure proposed in Figure 1C, this sample similarity information was used to generate a sample similarity (or dissimilarity) tree on which Equation 1 (adapted from [21]) was applied to assign a weight to each sample. After incorporating sample similarity information into weights, brain subregions were found to have lower weights compared with more distinctive tissues, such as testis and pituitary (Supplementary Figure S4, see Supplementary Data available online at http://bib.oxfordjournals.org/).

We proceeded to incorporate these weights to each of the specificity measures and compared the correlation of specificity values measured on the balanced and unbalanced sample sets to the correlations obtained before weights were applied. For this, each choice of a single brain subregion was used to generate a distinct balanced sample set ($n = 13$) that included a single brain subregion and all nonbrain samples; using these 13 replicates, we tested whether using the weighted specificity measure resulted in a stronger correlation between balanced and unbalanced sample sets. For all of the specificity measures tested, the correlation between the balanced and unbalanced sample sets increased when the weighting approach was applied compared with when weights were not applied. $P(R\ \text{Z-score}_{\text{weighted}} \leq R\ \text{Z-score}_{\text{flat}}) = 2.2e{-}16$; $P(R\ \text{Tau}_{\text{weighted}} \leq R\ \text{Tau}_{\text{flat}}) = 2.2e{-}16$; $P(R\ \text{Tsi}_{\text{weighted}} \leq R\ \text{Tsi}_{\text{flat}}) = 3.4e{-}10$; $P(R\ \text{Gini}_{\text{weighted}} \leq R\ \text{Gini}_{\text{flat}}) = 2.2e{-}16$; using the paired samples *t*-test for each test (Supplementary Figure S5, see Supplementary Data available online at http://bib.oxfordjournals.org/). Additionally, the improved correlation for the weighted measure held as the sample size varied while holding the proportion of the sample set composed of brain samples constant, except for Tsi which has previously been shown to be sensitive to sample size [11] (Supplemental Figure S6, see Supplementary Data available online at http://bib.oxfordjournals.org/). This trend of improved correlation between the balanced and unbalanced sample set was further replicated in the zebrafish and mouse single cell datasets. For the zebrafish dataset, the test results are as follows: $P(R\ \text{Z-score}_{\text{weighted}} \leq R\ \text{Z-score}_{\text{flat}}) = 2.9e{-}9$; $P(R\ \text{Tau}_{\text{weighted}} \leq R\ \text{Tau}_{\text{flat}}) = 5.2e{-}8$; $P(R\ \text{Tsi}_{\text{weighted}} \leq R\ \text{Tsi}_{\text{flat}}) = 3.3e{-}7$; $P(R\ \text{Gini}_{\text{weighted}} \leq R\ \text{Gini}_{\text{flat}}) = 2.6e{-}8$, using the paired samples *t*-test for each test. For the mouse dataset, the test results are as follows: $P(R\ \text{Z-score}_{\text{weighted}} \leq R\ \text{Z-score}_{\text{flat}}) = 3.7e{-}8$; $P(R\ \text{Tau}_{\text{weighted}} \leq R\ \text{Tau}_{\text{flat}}) = 5.7e{-}8$; $P(R\ \text{Tsi}_{\text{weighted}} \leq R\ \text{Tsi}_{\text{flat}}) = 8.2e{-}9$; $P(R\ \text{Gini}_{\text{weighted}} \leq R\ \text{Gini}_{\text{flat}}) = 1.4e{-}8$, using the paired samples *t*-test for each test (Supplemental Figures S7 and S8, see Supplementary Data available online at http://bib.oxfordjournals.org/).

When gene ontology enrichment analysis was performed on the weighted measures, the enrichment of brain related terms in the top 1% of genes with the largest positive difference in specificity measured between the balanced and unbalanced sample sets decreased for all measures, except for Tsi which did not change substantially (Supplemental Figure S9, see Supplementary Data available online at http://bib.oxfordjournals.org/). As was the case for the flat measures, for the weighted specificity measures, there was less consistency in terms associated with the genes where specificity in the balanced sample set was less than specificity in the unbalanced sample set across measures. However, for Tau, there was enrichment of brain related terms, possibly representing brain subregion specific genes being called as more specific as the sample set size increases with the inclusion of more brain subregions (Supplemental Figure S10, see Supplementary Data available online at http://bib.oxfordjournals.org/).

These results suggest that incorporating sample similarity information via weights allows one to include additional samples enriching the transcriptomic diversity within the sample set without necessarily sacrificing the ability to identify particular tissue- and cell-type-specific genes.

Validation of similarity-weighted specificity scores

To further test whether integration of similarity information improves the stability of gene specificity measures across variable sample sets, we used the GTEx dataset to quantify the degree of change in specificity scores as the proportion of the sample set composed of brain subregions increased for both the weighted and flat measures. The procedure to calculate the change in specificity is outlined in Figure 2A and B and a more detailed description of the procedure is included in Methods section.

Following the procedure outlined in Figure 2A and B, we observed that the weighted measures exhibited a marked reduction in the variance of the change in specificity measures as additional brain samples were added to the sample set (Figure 2C). When the similarity-weighting procedure was applied, specificity measurements were more stable as sampling depth into brain regions increased than when the procedure was not applied. For the nonbrain samples, assigning weights based on similarity to each tissue resulted in 73.1% (95% CI: 71.9–74.3%, paired samples *t*-test) lower variance in the change in specificity scores across all genes between the baseline with 1 brain sample included and the full set of 13 brain samples included than when weights were not used. For the brain samples, assigning weights based on similarity to each tissue resulted in 31.6% (95% CI: 25.4–37.3%, paired samples *t*-test) lower variance in the change in specificity scores across all genes between the baseline with 1 brain sample included and the full set of 13 brain samples included than when weights were not used (Figure 2C). A similar reduction in variance of specificity values between the baseline of 1 brain sample and inclusion of the full sample set was observed for Tau, Tsi and Gini as well (Figure 2C); however, for Tsi, this trend was not consistent at smaller sample sizes (Figure 2C, Supplemental Figure S6, see Supplementary Data available online at http://bib.oxfordjournals.org/). In contrast, when a similar procedure was used, substituting the brain partitioning with a random partitioning, we observed a much less dramatic difference in the change in variance between the weighted and flat measures. For the random partition, the nonbrain sample set was replaced by a random sample set of the same size, called P1, and the brain sample set was replaced with a random sample generated following the same incrementing procedure described in Figures 2A, called P2. We observed a <20% difference in the variance in specificity values in P1 and ~0% difference in the variance in specificity values in P2 between weighted and flat measures in the random partition compared with the 73.1% difference in P1 and the 31.6% difference in P2 between the weighted and flat measures in the brain partitioned sample set when the number of samples in P2 increased from 1 to 13 (Supplementary Figure S11, see Supplementary Data available online at http://bib.oxfordjournals.org/, Figure 2C).

As cut-off values are often used to binarize genes as either specific or nonspecific, we wanted to test whether incorporating sample similarity information would also improve the stability of gene sets called as specific as the sample set becomes more unbalanced. To do this, we compared the sets of genes that would be called as specific using different cutoff values as the number of brain subregions included in the sample set increased (Figure 2D). The Jaccard index, which is the ratio of the intersection and the union of two sets, was used to measure similarity of the gene sets. The Jaccard index ranges from 0, with no elements common to both sets, to 1, with all elements being shared between both sets. We observed that the set of genes specific to brain samples changed substantially over typical Z-score cutoff values between 2 and 3 SD. For example, at a Z-score cutoff of 2 SD, the Jaccard index dropped to 0.24 (95% CI: 0.14–0.34, *t*-test) for the flat measure, compared with a Jaccard index of 0.59 (95% CI: 0.54–0.63, *t*-test) at the same cutoff for the weighted measure as the number of brain samples included increased from 1 to 13 (Figure 2D). The change in the Jaccard index for the set of nonbrain sample specific genes was also substantial. At a Z-score cutoff of 2 SD, the Jaccard index dropped to 0.65 (95% CI: 0.64–0.67, *t*-test) for the flat measure, compared with a Jaccard index of 0.78 (95% CI: 0.77–0.79, *t*-test) at the same cutoff for the weighted measures as the number of brain samples included increased from 1 to 13 (Figure 2D). Similar but less dramatic trends were observed for Tau, Tsi and Gini measures (Figure 2D). In contrast, when the partition was random such that expanding the sample set included adding samples without high similarity to those already in the set, there was no significant difference in the change in Jaccard statistics between the weighted and flat measures (Supplementary Figure S12, see Supplementary Data available online at http://bib.oxfordjournals.org/).

Effect of integrating weights on patterns of specificity

We next wanted to explore the factors which influenced how a gene's specificity score changed in response to integration of weighted similarity information. We first looked at the GTEx dataset. The most striking change in specificity scores that occurred as the sampling depth of brain subregions increased were in genes with brain-related functions (Supplementary Figure S1, see

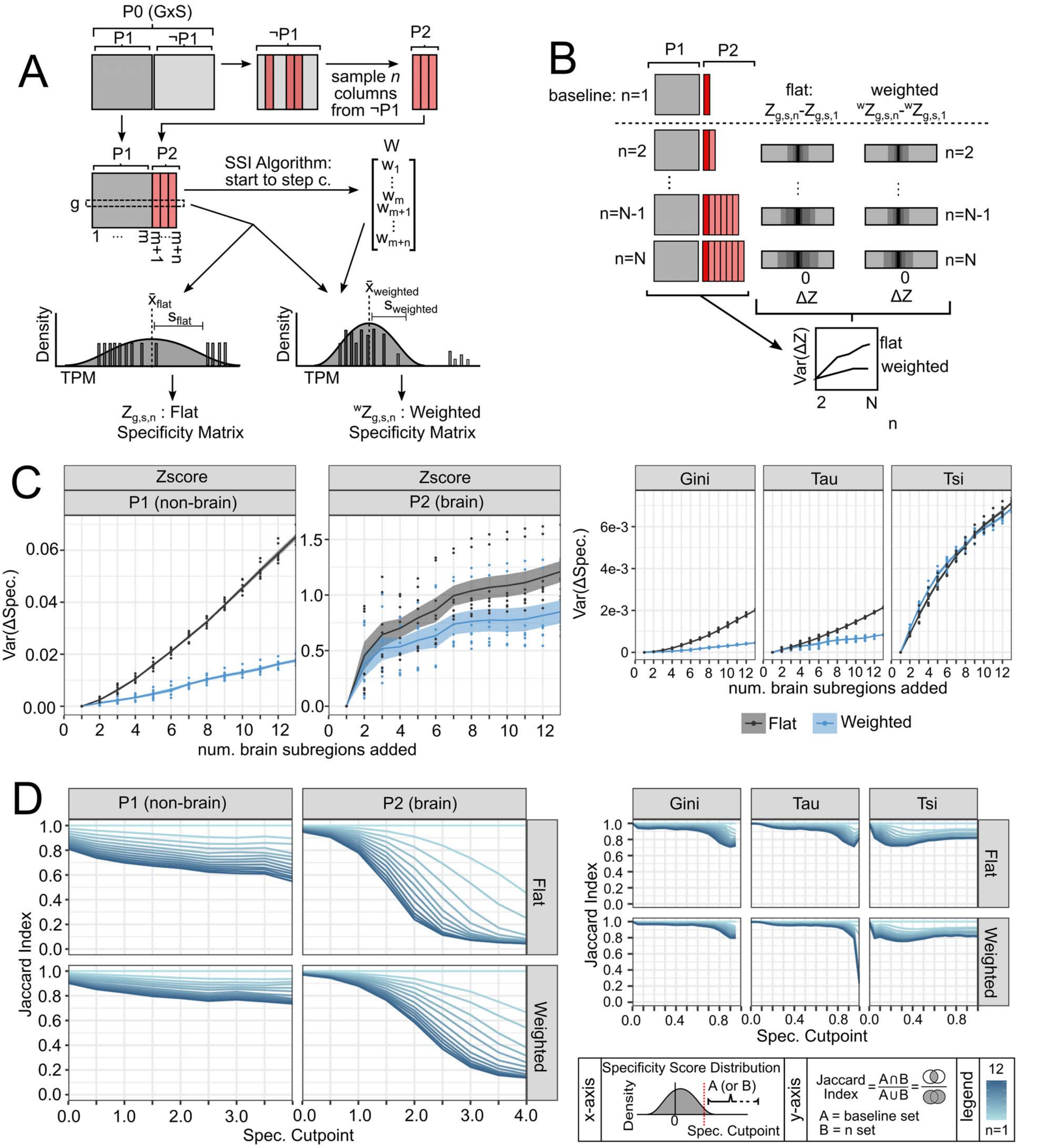

Figure 2. Quantification of robustness of specificity measures as sampling depth into brain subregions increases. (**A**) Workflow for generating each specificity matrix with the validation sampling procedure. P0 is the full dataset. G is the number of genes and S is the number of samples. P1 is the set of nonbrain origin tissue samples. P1 is the set of all brain origin tissue samples. P2 is a random selection of n brain samples. (**B**) Z-score is illustrated but a similar procedure was used for each specificity measure. Each specificity matrix ($Z_{g,s,n}$) where $n > 1$ is compared with a baseline where $n = 1$, where 1 brain sample was included in P2. Plotted on the x-axis is the density of the change in Z-scores (ΔZ) for all genes in all samples in either P1 or the brain sample initially selected as baseline from P2, with darker color indicating increased density. These data are finally summarized in the change in variance of the ΔZ values as the sample set increases to include more brain samples. Note: for P2, only the change in specificity scores associated with the brain sample selected for the baseline (darker red in figure) is recorded for **C** and **D** for each of 8 permutations of the procedure where each permutation involves selection of a different brain tissue sample for the baseline and a different ordering of the addition of the remaining brain samples to the sample set. (**C**) On the y-axis is the variance of change in specificity measure compared with a baseline dataset using 1 brain sample when *n* additional brain samples are added. The number of additional brain samples in P2 is given on the x-axis. For Z-score, specificity values associated with samples in P1 and P2 are plotted separately since Z-score can be associated with each tissue individually; other specificity measures aggregate across all samples so resolution of specificity between samples in P1 and P2 is not possible for these measures. Points are values from each of the 8 permutations of the procedure, lines are the mean values for each value of *n* and the shaded area is the 95% confidence interval (**D**). The x-axis is the cutoff above which a gene is called as specific. The y-axis is the jaccard index comparing overlap of the set of genes called as specific at the cutoff given on the x-axis

Supplementary Data available online at http://bib.oxfordjournals.org/). If incorporating sample similarity information reduced the bias introduced by increasing sampling depth in the brain, then we would expect that most of the differences between the weighted and flat specificity scores would occur in genes primarily expressed in the brain and with brain-related functions. Indeed, when we looked at the top 10 genes with the largest positive difference between the weighted and flat Z-score in each tissue (i.e., where the weighted specificity score was greater than the flat score), the genes with the largest change in specificity value were genes specific to brain samples (Figure 3A). Even in nonbrain tissues, the largest changes in gene specificity were in genes where expression was shared with brain samples (Figure 3A, Supplementary Figure S13, see Supplementary Data available online at http://bib.oxfordjournals.org/).

For the top 10 genes with the largest negative difference in specificity value between the weighted and flat Z-score in each tissue (i.e., where the weighted specificity score was less than the flat score), the largest effects were in nonbrain samples (Figure 3B). These changes in nonbrain samples tended to be in genes with high values of specificity from the flat measure being measured as slightly less specific by the weighted measure (Supplementary Figure S13, see Supplementary Data available online at http://bib.oxfordjournals.org/). This effect is likely due to the decrease in the effective sample size caused by downweighting individual brain samples. In the brain samples, while the effect size of the negative difference between the weighted and flat Z-score was modest (Figure 3B), this difference was associated with genes specifically depleted in brain samples becoming more specifically depleted, suggesting an increase in the power of the weighted Z-score to detect genes specifically depleted in brain tissues (Supplementary Figure S13, see Supplementary Data available online at http://bib.oxfordjournals.org/).

While we focused on protein coding genes for most of our analyses, we also looked at the specificity of lincRNA which have been observed to have strong patterns of tissue and cell-type specificity [22, 23]. As previously observed, we found that the proportion of lincRNAs with a high degree of tissue specificity was substantially greater than that for protein coding genes (Supplemental Figure S14, see Supplementary Data available online at http://bib.oxfordjournals.org/). The largest differences between the flat and weighted specificity values measured for lincRNAs were primarily in genes specific to brain samples (Supplemental Figures S14 and S15, see Supplementary Data available online at http://bib.oxfordjournals.org/), similar to what was observed for protein coding genes (Figure 3A, B, Supplemental Figures S13 and S14, see Supplementary Data available online at http://bib.oxfordjournals.org/).

We next looked at the behavior of the flat and weighted specificity scores for genes known to have tissue-specific expression patterns. Brain-specific genes *OLIG1* and *OLIG2* are markers for oligodendrocytes, cells which are restricted to the spinal cord and brain [24] and which are somewhat less abundant in the cerebellum than other brain regions [25]. With the flat Z-score, *OLIG1* had specificity values in brain samples between 1.02–1.06 SDs in the cerebellar subregions and 1.56–2.03 SDs in the other brain regions, and *OLIG2* had specificity values between 0.91–0.97 SDs in the cerebellar subregions and 1.56–2.11 SDs in the other brain regions. When measured using the weighted Z-score, *OLIG1* had specificity values between 1.54–1.60 in cerebellar subregions and 2.22–2.75 SDs in other brain regions, and *OLIG2* had specificity values between 1.49–1.57 SDs in cerebellar subregions and 2.33–3.03 SDs in other brain regions for the weighted Z-score. Specificity estimates for the more specific basal ganglia marker, *DRD1* [26], also increased when weights were applied, up to 4.77–5.00 SDs in brain basal ganglia subregions from the 3.22–3.30 SDs by the flat Z-score (Figure 2D, Supplemental Table S1, see Supplementary Data available online at http://bib.oxfordjournals.org/). The specificity scores between the flat and weighted measures were similar for genes specific to uniquely represented tissue types. For example, *PRSS2*, a pancreas-specific protease, had a flat Z-score of 6.38 SDs and a weighted Z-score of 5.40 SDs in the pancreas, and *MYH6*, a heart-specific myosin heavy chain, had a flat Z-score of 5.16 SDs and a weighted Z-score of 4.77 in the atrial appendage of the heart (Figure 3C). Similar trends for these marker genes were observed for Tau, Tsi and Gini coefficients (Supplemental Table S1, see Supplementary Data available online at http://bib.oxfordjournals.org/). Further quantification of the differences in genes called as specific between the flat and weighted measure showed a general increase in the number of genes called as specific to brain tissues when weights were applied and modest differences in the set of genes called as specific between the flat and weighted measures (Supplemental Figure S16, see Supplementary Data available online at http://bib.oxfordjournals.org/).

We next performed gene ontology enrichment analysis on the genes that changed from nonspecific to specific in either direction using a Z-score cutoff for classification as specific of 2 SDs. The top 10 terms were those related to synaptic and neurotransmitter function (e.g. synapse organization, neurotransmitter secretion, signal release from synapse) (Figure 3D). This is consistent with the expectation that weighting based on sample similarity

relative to the baseline set where P2 includes only 1 randomly selected brain region sample. For Z-score, the jaccard index is the average over all samples in the sample set (P1 or P2), for all other measures which aggregate across all samples the jaccard index is obtained directly. Line color corresponds to the value of *n*. Note: at *n* = 1, the sample set used is the same as the sample set used in the baseline resulting in the line at jaccard index = 1 for *n* = 1 in all cases.

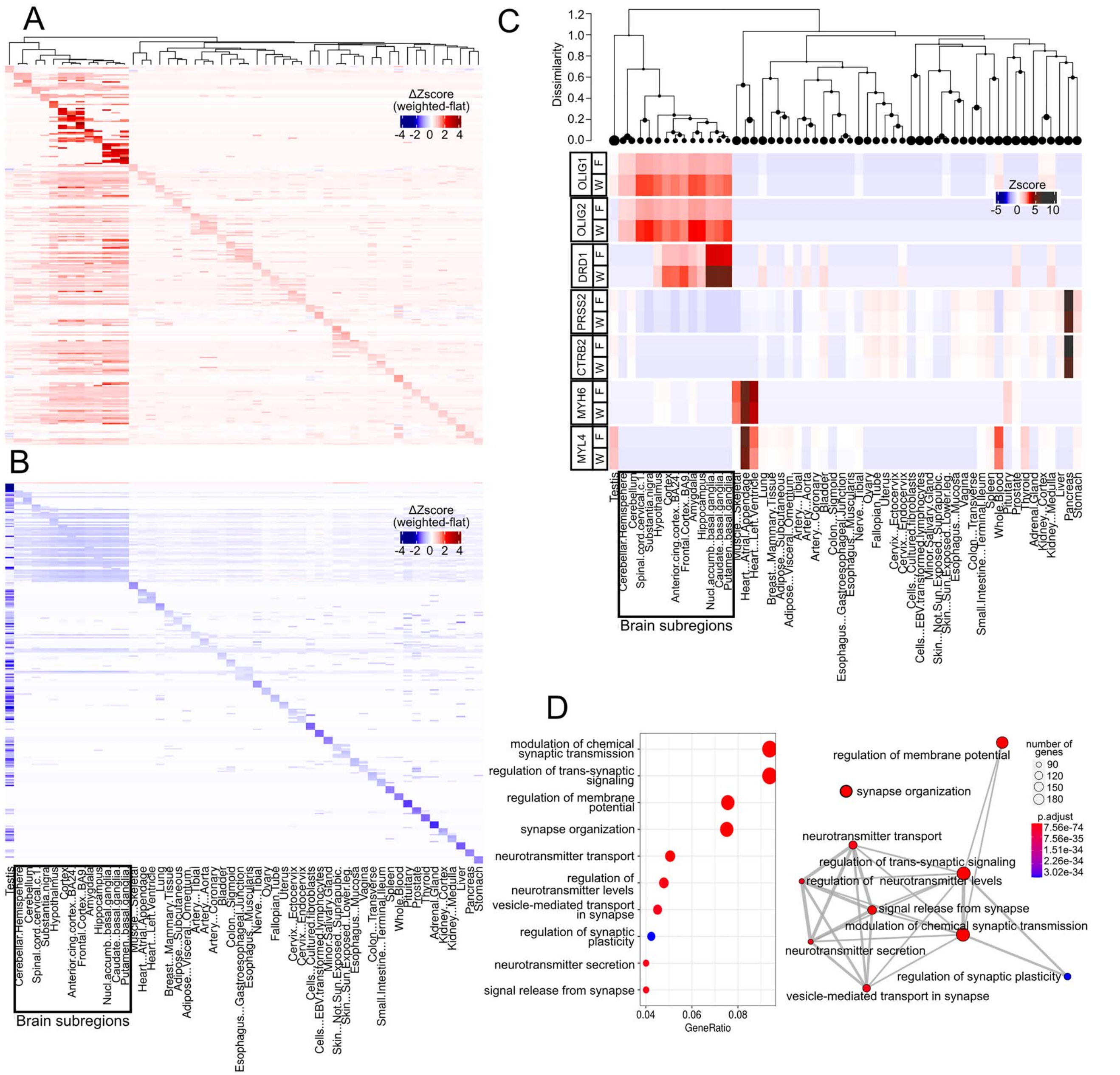

Figure 3. Biological context of differences between flat and weighted measures of specificity. (**A**) Top 10 genes with greatest difference between weighted and flat Z-score for each tissue. Rows correspond to individual genes, columns to tissues. Note: diagonal produced by having top 10 genes from leftmost tissue on x-axis as first 10 rows, top 10 from next leftmost tissue as next 10 rows, etc. (**B**) Bottom 10 genes with greatest difference between weighted and flat Z-score for each tissue (**C**) genes known to be specific to brain regions, *OLIG1*, *OLIG2* and *DRD1*, pancreas, *PRSS1* and *CTRB2*, and heart *MYH6* and *MYL4*. For each gene, top row is flat (F) Z-score, bottom row is weighted (W) Z-score. Dendrogram at top shows the dissimilarity tree used to generate sample weights which are shown as the area of the leaf nodes of the dendrogram. (**D**) Gene ontology results highlighting top 10 terms in the set of genes that have specificity values <2 SD by the flat Z-score, and >2 SD by the weighted Z-score. On the left is the Gene Ratio, i.e. the proportion of the gene set with the given GO term. On the right is the network plot, where the edge width indicates the number of shared genes between a connected pair of terms.

would increase power to detect genes that are specific to tissues that are more deeply sampled and overrepresented in the sample set.

We next repeated these analyses for the zebrafish and mouse single cell datasets. In zebrafish, there was an overrepresentation of cell clusters from brain-related cell types as well as a secondary overrepresentation of cell clusters from skeletal muscle cell types. When looking at the top 5 genes from each cell cluster with the greatest positive difference in specificity value measure between the weighted and flat specificity scores, the genes specific to brain and skeletal muscle clusters had the largest absolute change in specificity values measures (Supplemental Figure S17A, see Supplementary Data available online at http://bib.oxfordjournals.org/). Those genes with the largest change in other cell clusters tended to be those with expression shared with brain or skeletal muscle (Supplemental Figure S17A, see Supplementary Data available online at http://bib.oxfordjournals.org/). For the bottom 5 genes with the largest negative difference in specificity between the weighted and flat specificity scores, the genes with the largest absolute change in specificity value measures were those with high specificity to nonbrain and nonskeletal muscle cell clusters being measured as slightly less specific (Supplemental Figures S17B and S18, see Supplementary Data available online at http://bib.oxfordjournals.org/), likely due to a decrease in the effective sample size between the flat and weighted measures. In the brain and skeletal muscle cell clusters, the largest negative change in specificity values occurred in genes depleted in brain and skeletal muscle cell clusters being measured as more specifically depleted (Supplemental Figures S17B and S18, see Supplementary Data available online at http://bib.oxfordjournals.org/) suggesting an increase in the power to detect specifically depleted genes in these overrepresented cell types when using the weighted specificity measure. In the mouse, the same patterns were observed where genes specific to the overrepresented myeloid lineage cell clusters and the kidney cell clusters were those that had the greatest positive difference between the weighted and flat measures (Supplemental Figure S19A, see Supplementary Data available online at http://bib.oxfordjournals.org/), and those with the more modest negative difference overlapped more with genes specifically depleted in the myeloid and kidney cell types being measured as more specifically depleted (Supplemental Figures S19B and S20, see Supplementary Data available online at http://bib.oxfordjournals.org/).

As observed in the GTEx dataset, when we looked at markers genes in the zebrafish and mouse cell types that were either uniquely represented or overrepresented, we found similar levels of specificity between the weighted and flat specificity measures for those cell types that were uniquely represented and an increase in specificity for markers for overrepresented cell types (Supplemental Figures S17C and S19C, see Supplementary Data available online at http://bib.oxfordjournals.org/). In the zebrafish dataset, gene ontology enrichment analysis of the terms associated with genes that were called as nonspecific with the flat measure and specific with the weighted measure found enrichment in terms related to the overrepresented brain and skeletal muscle cell types (e.g., muscle cell development, muscle contraction, brain development, head development) (Supplemental Figure S17D, see Supplementary Data available online at http://bib.oxfordjournals.org/). In the mouse datasets, a similar trend was observed with the top terms being those related to the overrepresented myeloid cell types; however, no terms related to the other overrepresented cell type, kidney cells, were observed in the top 15 enriched terms (Supplemental Figure S19D, see Supplementary Data available online at http://bib.oxfordjournals.org/).

Overall, these results demonstrate that the use of our sample similarity weighting procedure improves the stability of gene specificity measures across a variety of sample sets that are balanced or unbalanced with particular tissue or cell types overrepresented. This enables the identification of genes specific to more deeply sampled biological contexts and reduces bias that is otherwise introduced by variation in sampling depth. Implementing this weighting procedure can give researchers more flexibility in building a sample set, allowing greater sampling depth into a cell type, tissue or organ of interest without sacrificing the ability to detect genes specific to that same cell type, tissue or organ.

Discussion

Previous work developing and implementing measures of specificity have had a variety of aims including imputation of expression levels for cell and tissue precursors [14], investigating mechanisms of dosage compensation [15] and characterizing conservation of gene expression patterns across evolutionary time [27, 28]. While existing measures have been used successfully, we identified a limitation in that these measures lack a mechanism to account for the similarities that exist between cells or tissues. The absence of a mechanism to account for sample similarity makes existing specificity measures sensitive to the choice of sample set used and can introduce bias into analyses, an issue that has been previously noted [11, 29]. A feature of this sensitivity to the sample set composition is a loss of measure robustness as the sampling depth of particular developmental lineages increases, particularly for the features that are specific to the more deeply sampled lineage. Greater depth of sampling is necessary for a more complete view of transcriptome diversity, and therefore, the antagonistic relationship between sampling depth and the stability of specificity measures is problematic.

To address this, we utilized sample similarity information to weight each sample's contribution to measures of gene specificity. In this work, we have shown that accounting for similarity between biological samples in the manner proposed makes measures of specificity more robust to sample set variation and improves the ability of these measures to detect features specific to different cell and tissue types, even when the cell or tissue type is overrepresented within the larger sample set.

One component of the procedure proposed here for integrating sample similarity information into measures of gene specificity is the use of a similarity (or dissimilarity) tree structure to partition weight across the sample set, analogous to the method for assigning sequence weight used by the multiple sequence alignment algorithm *ClustalW* [21]. This mechanism is a natural choice when samples can be defined along a natural hierarchy, such as when the developmental relation between a set of cells is known; however, for tissues, which are often composites of cells from distinct lineages, this model is imprecise. While we have demonstrated that using this model to weight samples improves existing measures of gene specificity for tissues, more general graph-based methods that can account for heterogeneous tissue composition may be able to improve upon the method proposed here by refining the weighting of samples for heterogeneous samples.

Applying this workflow on single-cell data avoids the issue of dealing with heterogeneous composites and also provides a higher resolution view of patterns of specificity for gene expression. However, single-cell analysis requires dealing with problems of low read depth and accurate transcript estimation among others [30]. Furthermore, as the method proposed here involves calculating a similarity matrix between samples which requires $O(n^2)$ time, performing the calculation on a large dataset of tens of thousands or more cells becomes, though feasible, somewhat resource intensive without additional optimizations. Clustering cells is a common part of most workflows for single cell analysis and provides a convenient work around for these issues [31]. Here, we have shown that the SSI procedure can be used on clusters of single cells to achieve improvements to specificity estimates within single cell analyses.

As additional RNA-seq datasets come online, particularly those spanning various stages of development, our method for calculating specificity that is robust to expansion of the sample set will be invaluable. The Developmental Genotype-Tissue Expression (dGTEx) project has recently been announced and will expand on the GTEx project to include samples from neonatal, pediatric and adolescent individuals. dGTEx will add depth to a large range of developmental stages for many cells, tissues and organs and will provide a unique opportunity to broadly investigate transcriptomic changes through development [32]. The method for calculating gene specificity proposed here is a natural model for hierarchical developmental relationships that will be captured in this dataset and that currently exist in datasets for model organisms [19, 20, 33–35]. We expect that our method can be used to facilitate improved investigations into the dynamics of gene expression across development in a transcriptome-wide context.

Here, we have demonstrated that integrating sample similarity information into measures of gene expression specificity in cells and tissues improves the robustness of these measures to variation in the underlying sample set. By improving the stability of specificity measures to deeper sampling of particular biological contexts of interest, the proposed procedure can facilitate the analysis of patterns of gene expression that captures both the broad, by including a diverse set of cell or tissue types, as well as the focused perspective, by allowing greater depth of sampling of highly similar cell or tissue types. This procedure for integrating sample similarity can easily be extended to measure the specificity of other functional measures of the genome and epigenome such as histone modification or DNA methylation features.

Methods

Data availability

The GTEx data used for the analyses described in this manuscript were obtained from the Genotype-Tissue Expression (GTEx) Project which was supported by the Common Fund of the Office of the Director of the National Institutes of Health and by NCI, NHGRI, NHLBI, NIDA, NIMH and NINDS [18].

website: https://www.gtexportal.org/home/datasets.

access date: 1 March 2022.

file: GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_median_tpm.gct.gz.

The zebrafish single cell dataset comes from [19].

website: https://cells.ucsc.edu/zebrafish-dev.

access date: 1 March 2022.

cell annotation file: meta.tsv.

cell expression file: exprMatrix.tsv.gz.

The mouse single cell dataset comes from [20].

website: https://figshare.com/articles/dataset/HCL_DGE_Data/7235471.

access date: 1 March 2022.

cell annotation file: annotation_rmbatch_data_revised417.zip.

cell expression file: dge_rmbatch_data.tar.gz.

Data preprocessing

For the data from GTEx, transcripts per million (TPM) values were recalculated after removing mitochondrial gene reads, to prevent signal driven by relative mitochondrial abundance in tissues, and after removing nonprotein coding genes. Expression values in TPM were then log transformed as log10(TPM + 1). The addition of 1 to TPM value before taking the log was done to avoid the issue of taking log of 0, and also because very low TPM estimates are unstable across replicates at standard sequencing depths in the tens of millions of reads. Following this transformation, log10(TPM +1) values for gene expression were scaled with median normalization across all samples [36].

For the single cell data from the zebrafish and mouse datasets, cell cluster annotations were obtained from their respective source studies. These cluster annotations can be found in their respective ‘cell annotation file’ linked in Data availability section above. RNA read counts were obtained from their respective ‘cell expression file’ and reads were aggregated across all cells within each cluster. Reads from mitochondrial and noncoding RNAs were filtered out. Clusters with less than 100 k total reads after this filtering were then removed from further analysis. Genes with read counts <10 for each cluster were set to 0 to reduce noise caused by low read counts. Read counts for each gene were then transformed to TPM values by multiplying read counts by 1e6 and dividing by the sum of read counts for each cluster. These TPM values were then log10(TPM + 1) transformed. These transformed log10(TPM + 1) values were scaled with median normalization across all samples [36].

General algorithm for incorporating sample similarity information into measures of gene specificity

The SSI Algorithm in Figure 1C outlines the general workflow for integrating sample similarity information with an arbitrary measure of specificity. Beginning with a matrix of log transformed gene expression values for a set of samples (genes as rows, samples as columns) sample similarity is measured (SSI step a.). The use of the gene expression matrix for measuring sample similarity is suggested as the gene expression matrix is already required for measuring gene specificity; however, other feature sets could be used to assign sample similarity. The important component is to have a mechanism for generating a meaningful sample similarity matrix. Several measures of similarity (cosine, canberra, euclidean, manhattan) were tested and each of the similarity measures tested produced similar intuitive sample similarity structure. For example, each measure found brain samples to have a high degree of similarity with one another. The major difference in measures of similarity was the average similarity across all pairs of samples (Supplemental Figure S21, see Supplementary Data available online at http://bib.oxfordjournals.org/). For downstream analyses, cosine similarity was used as it has previously been shown to be robust in high-dimensional datasets in benchmarking studies [37, 38]. The next step is to apply a hierarchical clustering algorithm on the sample similarity matrix (SSI step b.). Single, average and complete clustering were tested and each produced similar intuitive clusters of samples (e.g., brain samples clustered together; tibial, aortic and coronary arteries clustering together; etc.) (Supplemental Figure S22, see Supplementary Data available online at http://bib.oxfordjournals.org/). Average linkage clustering was used as it has previously been shown to be robust when the size of cluster groups varies substantially [39]. Other methods could be substituted so long as a suitable tree structure is generated for sample representation, where suitability can be determined, for example, on metrics such known developmental relations between tissues or cells. The dissimilarity tree is then used to determine the sample weights (SSI step c.) with the recursive function given in Eq. 1 and described in the section below. The final step is to use a specificity function that allows sample weights with the initial log transformed expression value matrix (SSI step d.). The specificity functions used in this paper are discussed below.

Assignment of sample weights

Sample weights are assigned using the recursive function

$$w_i = \frac{d_{i,p(i)}}{n_i} + w_{p(i)} \tag{1}$$

,where w_i is the weight of node i in the dissimilarity tree (where dissimilarity = 1 − similarity). $p(i)$ is the parent of node i. $d_{i,p(i)}$ is the distance between node i and its parent node $p(i)$. n_i is the number of descendant leaf nodes for node i, where a leaf node is considered a descendant of itself. Weight of the root node is set to zero. Weighting method is based on that introduced for the guide tree implemented in the ClustalW sequence alignment algorithm [21]. Supplemental Figure S23 (see Supplementary Data available online at http://bib.oxfordjournals.org/) provides an example calculation.

Specificity measures tested

Four different specificity scores were used to measure how changes in the depth of sampling of certain regions affected the variance in specificity scores assigned to genes. For each equation, n is the number of tissues and x_i is the expression of a gene of interest in tissue i.

The first measure is Z-score [13], which determines specificity by calculating how many SD away gene expression in a given tissue is from the mean expression value across all tissues for that gene. It is calculated as

$$Z_i = \frac{x_i - \bar{x}}{s}, \tag{2}$$

where

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{2.1}$$

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}, \tag{2.2}$$

where Z_i is the Z-score in tissue i, x_i is the gene expression value in tissue i, $\bar{x}$ is mean expression of the gene of interest across all tissues and s is the SD in expression of the gene of interest across tissues. The more positive the Z-score, the more specific a certain gene is to a certain tissue.

The weighted version of this equation is given by

$$Z_{wi} = \frac{x_i - \bar{x}_w}{s_w}, \tag{3}$$

where, from [40],

$$\bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \tag{3.1}$$

$$s_w = \sqrt{\frac{\sum_{i=1}^{n} w_i \left(x_i - \overline{x}_w\right)^2}{\sum_{i=1}^{n} w_i}} \tag{3.2}$$

and w_i is the weight of a given tissue, and other variables are the same as in Eqs 2, 2.1–2.2.

The second measure is tau (τ) [14], which is a tissue specificity measure ranging from 0 to 1, with genes with tau near 0 being more ubiquitously expressed and scores near 1 being more specifically expressed. At the extremes, a score of 0 corresponds to a gene with equal expression across all tissues, while a score of 1 represents a gene only expressed in one tissue. In a benchmark of measures for gene specificity, tau was found to be consistently the most robust measure of gene specificity on several metrics [11]. Tau is calculated as

$$\tau = \frac{\sum_{i=1}^{n} \left(1 - \hat{x}_i\right)}{n-1}, \tag{4}$$

where

$$\hat{x}_i = \frac{x_i}{\max\limits_{j\in\{1,\ldots,n\}} x_j} \tag{4.1}$$

with the weighted version of the equation being

$$\tau_{weighted} = \frac{\sum_{i=1}^{n} \left(w_i - w_i\hat{x}_i\right)}{\left(\sum_{i=1}^{n} w_i\right) - w_k} \tag{5}$$

with

$$k \text{ such that } x_k = \max_{j\in\{1,\ldots,n\}} x_j, \tag{5.1}$$

where $\hat{x}_i$ is the same as that for Eq. 4 given in Eq. 4.1. The range, which we define as the set of possible output values, of the weighted tau is the same as the unweighted tau.

The third measure is tissue specificity index (Tsi) [15], which measures specificity on a scale of $1/n$ to 1. For any given gene, $1/n$ represents equal gene expression across tissues, while 1 represents expression only in one tissue. Tsi is calculated as

$$tsi = \frac{\max\limits_{j\in\{1,\ldots,n\}} x_j}{\sum_{i=1}^{n} x_i} \tag{6}$$

with the weighted version of the equation being

$$tsi_{weighted} = \frac{w_{\arg\max_{j\in\{1,..,n\}} x_j} \max\limits_{j\in\{1,\ldots,n\}} x_j}{\sum_{i=1}^{n} w_i x_i} \tag{7}$$

.The weighted Tsi has a similar range as the unweighted version, except that the lower bound is $w_k / \sum_{i=1}^{n} w_i$ with k such that $x_k = \max\limits_{j\in\{1,\ldots,n\}} x_j$, instead of $1/n$.

The fourth specificity measure was the Gini coefficient [16, 17], a measure of inequality commonly used in economics. Existing on a 0 to $(n-1)/n$ scale, for any gene of interest, a score of 0 represents uniform distribution of gene expression across tissues, while a score of $(n-1)/n$ would indicate that a gene is only expressed in one tissue. The Gini coefficient is calculated as

$$Gini = \frac{n+1}{n} - \frac{2\sum_{i=1}^{n} (n+1-i)\, x_i}{n\sum_{i=1}^{n} x_i}, \tag{8}$$

where x_i are ordered from least to greatest.

The weighted version from [41] is given by

$$Gini_{weighted} = 2\sum_{i=1}^{n} w_i \left(x_i - \overline{x}\right) \left(\hat{F}_i - \overline{F}\right) / \overline{x}, \tag{9p}$$

where

$$\hat{F}_i(x) = \sum_{j=0}^{i-1} w_j + w_i/2 \tag{9.1}$$

with $w_0 = 0$ and again with x_i ordered from least to greatest $\overline{F}$ is the mean of $\hat{F}_i$. The range of the weighted Gini index is similar to the unweighted version except that the upper bound is $\left(\left(\sum_{i=1}^{n} w_i\right) - 1\right) / \sum_{i=1}^{n} w_i$ instead of $(n-1)/n$.

Specificity measure robustness testing

The GTEx dataset was used for the specificity measure robustness testing. To test the robustness of measures of specificity, the change in specificity estimates as the dataset came to contain an increasing proportion of brain samples was followed. For this, the GTEx dataset was used, which consists of 54 tissue types in total, of which 13 (25%) are from different brain subregions. The GTEx dataset was partitioned into 41 nonbrain tissues, P1, and 13 brain tissues, ¬P1. The following procedure was then repeated 8 times using a unique brain subregion sample for the baseline and a unique order of addition for the remaining brain subregion samples.

To begin, one brain sample was selected at random and placed in P2, this is $P2_{baseline}$. The union of P1 and $P2_{baseline}$, P1 U $P2_{baseline}$, was then taken as the sample set. Specificity was then measured using the P1 U $P2_{baseline}$ sample set with each of the flat and weighted specificity measures. The results generated using a single randomly selected brain sample serve as the baseline to compare estimates of specificity as additional brain samples were added to the sample set.

Next brain samples were added successively to $P2_{baseline}$ and specificity recalculated on P1 U $P2_{n=i}$, where *i* is the number of brain samples in P2 in the current iteration. The variance in the change in specificity between specificity measured on P1 U $P2_{n=i}$ and P1 U $P2_{\text{baseline}}$ across all genes was recorded and used in generating Figure 2C. The sets of genes called as specific at various cutoff values from the specificity values measured on P1 U $P2_{n=i}$ and on P1 U $P2_{\text{baseline}}$ were compared using the Jaccard index. The Jaccard index was recorded and used in generating Figure 2D. This was repeated until the sample set included all 13 brain tissues.

Gene ontology analysis

The clusterProfiler package [42] in R was used to perform enrichment analyses and generate gene ontology [43] plots. Sets of genes were defined as specified in relevant sections of text or figure captions and enrichment was tested against the set of all genes in the GTEx, mouse or zebrafish expression matrix after filtering nonprotein coding and mitochondrial genes. Benjamini–Hochberg procedure [44] was used to adjust *P*-values for significance. The Biological Process set of GO terms was used throughout.

Key Points

- Existing measures of gene specificity exhibit bias against genes specific to biological contexts that are overrepresented in the sample set.
- Adjusting sample weight based on sample similarity improves the stability of specificity measures even in sample sets where a specific biological context is overrepresented.
- The proposed workflow enables greater flexibility in the choice of the sample sets used for measuring specificity of gene expression.

Supplementary data

Supplementary data are available online at http://bib.oxfordjournals.org/.

Code availability

All code used for analyses in this manuscript are available at: https://github.com/leroybondhus/gene_specificity.

Authors’ Contributions

L.B. and V.A. developed the study. L.B., R.V. and Y.H. analyzed the data and wrote the manuscript with significant input from V.A.

Acknowledgements

We thank the authors and contributors to the Gene-Tissue Expression project for creating this open resource. We thank the members of the Arboleda lab for their constructive comments on the development of this project.

Funding

National Institutes of Health (DP5OD024579 to V.A., T32HG002536GATP to L.B., T32LM012424 to L.B.).

References

1. Vickaryous MK, Hall BK. Human cell type diversity, evolution, development, and classification with special reference to cells derived from the neural crest. *Biol Rev Camb Philos Soc* 2006;**81**(3): 425–55.
2. Barshir R, Shwartz O, Smoly IY, *et al.* Comparative analysis of human tissue Interactomes reveals factors leading to tissue-specific manifestation of hereditary diseases. *PLoS Comput Biol* 2014;**10**(6):e1003632.
3. Hekselman I, Yeger-Lotem E. Mechanisms of tissue and cell-type specificity in heritable traits and diseases. *Nat Rev Genet* 2020;**21**(3):137–50.
4. Lage K, Hansen NT, Olof Karlberg E, *et al.* A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. *Proc Natl Acad Sci U S A* 2008;**105**(52): 20870–5.
5. Cunha DL, Arno G, Corton M, *et al.* The Spectrum of PAX6 mutations and genotype-phenotype correlations in the eye. *Gen* 2019;**10**(12):1050. https://doi.org/10.3390/genes10121050.
6. Arboleda VA, Fleming A, Barseghyan H, *et al.* Regulation of sex determination in mice by a non-coding genomic region. *Genetics* 2014;**197**(3):885–97.
7. Genuth NR, Barna M. Heterogeneity and specialized functions of translation machinery: from genes to organisms. *Nat Rev Genet* 2018;**19**(7):431–52.
8. Herrmann H. Mechanisms of cell specialization. *Invest Ophthalmol* 1969;**8**(1):17–25.
9. Saitou M, Gaylord EA, Erica X, *et al.* Functional specialization of human salivary glands and origins of proteins intrinsic to human saliva. *Cell Rep* 2020;**33**(7):108402.
10. Conesa A, Madrigal P, Tarazona S, *et al.* A survey of best practices for RNA-Seq data analysis. *Genome Biol* 2016;**17**(January):13.
11. Kryuchkova-Mostacci N, Robinson-Rechavi M. A benchmark of gene expression tissue-specificity metrics. *Brief Bioinform* 2017;**18**(2):205–14.
12. GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, *et al.* Genetic effects on gene expression across human tissues. *Nature* 2017;**550**(7675):204–13.
13. Vandenbon A, Nakai K. Modeling tissue-specific structural patterns in human and mouse promoters. *Nucleic Acids Res* 2010;**38**(1):17–25.
14. Yanai I, Benjamin H, Shmoish M, *et al.* Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* 2005;**21**(5):650–9.
15. Julien P, Brawand D, Soumillon M, *et al.* Mechanisms and evolutionary patterns of mammalian and avian dosage compensation. *PLoS Biol* 2012;**10**(5):e1001328.
16. Ceriani L, Verme P. The origins of the Gini index: extracts from Variabilità E Mutabilità (1912) by Corrado Gini. *J Econ Inequal* 2012;**10**(3):421–43.
17. Gini, C. 1912. *Variabilità E Mutabilità*. ui.adsabs.harvard.edu.
18. GTEx Consortium. The genotype-tissue expression (GTEx) project. *Nat Genet* 2013;**45**(6):580–5.

19. Farnsworth DR, Saunders LM, Miller AC. A single-cell transcriptome atlas for zebrafish development. *Dev Biol* 2020;**459**(2): 100–8.
20. Han X, Wang R, Zhou Y, *et al.* Mapping the mouse cell atlas by microwell-Seq. *Cell* 2018;**172**(5):1091–1107.e17.
21. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 1994;**22**(22):4673–80.
22. Cabili MN, Trapnell C, Goff L, *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* 2011;**25**(18):1915–27.
23. Gloss BS, Dinger ME. The specificity of long noncoding RNA expression. *Biochim Biophys Acta* 2016;**1859**(1):16–22.
24. Miller RH. Regulation of oligodendrocyte development in the vertebrate CNS. *Prog Neurobiol* 2002;**67**(6):451–67.
25. Valério-Gomes B, Guimarães DM, Szczupak D, *et al.* The absolute number of oligodendrocytes in the adult mouse brain. *Front Neuroanat* 2018;**12**(October):90.
26. Cadet JL, Jayanthi S, McCoy MT, *et al.* Dopamine D1 receptors, regulation of gene expression in the brain, and neurodegeneration. *CNS Neurol Disord Drug Targets* 2010;**9**(5):526–38.
27. Assis R, Bachtrog D. Neofunctionalization of young duplicate genes in drosophila. *Proc Natl Acad Sci U S A* 2013;**110**(43): 17409–14.
28. Piasecka B, Robinson-Rechavi M, Bergmann S. Correcting for the bias due to expression specificity improves the estimation of constrained evolution of expression between mouse and human. *Bioinformatics* 2012;**28**(14):1865–72.
29. Martínez O, Humberto Reyes-Valdés M. Defining diversity, specialization, and gene specificity in transcriptomes through information theory. *Proc Natl Acad Sci U S A* 2008;**105**(28):9709–14.
30. Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp Mol Med* 2018;**50**(8):1–14.
31. Duò A, Robinson MD, Soneson C. A systematic performance evaluation of clustering methods for single-cell RNA-Seq data. *F1000Research* 2018;**7**(July):1141.
32. National Advisory Council for Human Genome Research (NACHGR). 2020. "Concept Clearance for FOAs Developmental Genotype-Tissue Expression (dGTEx)." https://www.genome.gov/sites/default/files/media/files/2020-02/Concept_Document_Developmental_GTEx.pdf (10 February 2020, date last accessed).
33. Cao J, Packer JS, Ramani V, *et al.* Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* 2017;**357**(6352):661–7.
34. Leader DP, Krause SA, Pandit A, *et al.* FlyAtlas 2: a new version of the drosophila melanogaster expression atlas with RNA-Seq, miRNA-Seq and sex-specific data. *Nucleic Acids Res* 2018;**46**(D1):D809–15.
35. Smith CM, Hayamizu TF, Finger JH, *et al.* The mouse gene expression database (GXD): 2019 update. *Nucleic Acids Res* 2019;**47**(D1):D774–9.
36. Dillies M-A, Rau A, Aubert J, *et al.* A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief Bioinform* 2013;**14**(6):671–83.
37. Deshpande R, Vandersluis B, Myers CL. Comparison of profile similarity measures for genetic interaction networks. *PLoS One* 2013;**8**(7):e68664.
38. Shirkhorshidi AS, Aghabozorgi S, Wah TY. A comparison study on similarity and dissimilarity measures in clustering continuous data. *PLoS One* 2015;**10**(12):e0144059.
39. Ferreira L, Hitchcock DB. A comparison of hierarchical methods for clustering functional data. *Commun Stat Simul Comput* 2009;**38**(9):1925–49.
40. Price GR. Extension of covariance selection mathematics. *Ann Hum Genet* 1972;**35**(4):485–90.
41. Lerman RI, Yitzhaki S. Improving the accuracy of estimates of Gini coefficients. *J Econom* 1989;**42**(1):43–7.
42. Yu G, Wang L-G, Han Y, *et al.* clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics* 2012;**16**(5):284–7.
43. Ashburner M, Ball CA, Blake JA, *et al.* Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat Genet* 2000;**25**(1):25–9.
44. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodology* 1995;**57**(1):289–300.