Predicting gene-function associations using tissue-specific gene expression

Junjie Zhu, Nico Chaves, Noam Weinberger
Mentors: Rok Sosič, Marinka Zitnik
June 7, 2016

1. Problem description

The variation of gene expression in different human tissues leads to the functional diversity across these tissues. Here we address whether gene expression in certain tissues is informative of biological processes in other tissues. We frame this problem as a gene-function prediction problem. In particular, we address whether tissue-specific gene expression can be used to predict associations between genes and biological functions in other tissues.

2. Data

We used the Genotype Tissue Expression Consortium (GTEx) dataset [1] to obtain tissue-specific RNA-seq gene expression profiles. The dataset consists of samples taken from various tissues of 544 individuals. GTEx contains 8,555 samples in total, drawn from 53 different tissues. For each sample, it provides expression levels measured for approximately 192,000 RNA transcripts. We also used gene ontology (GO) [2] to obtain known associations between genes and biological processes, focusing only on those processes that have been annotated as tissue-specific [3].

2.1 Preprocessing

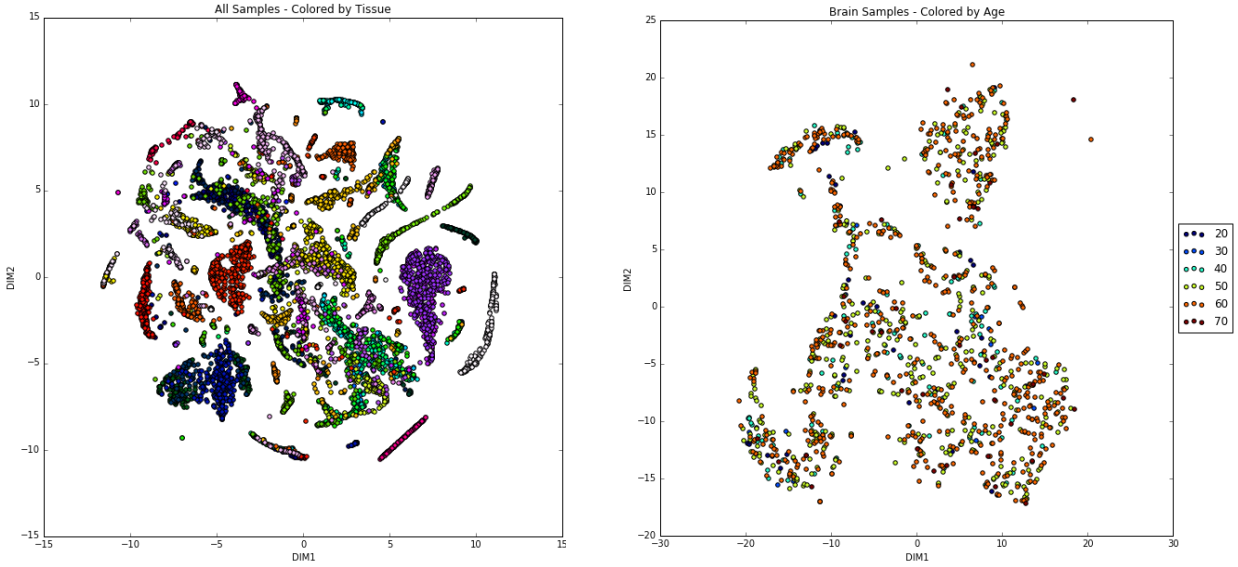
We downloaded the GTEx data as a 192,000 x 8,555 matrix, in which rows correspond to RNA transcripts, columns correspond to samples, and matrix elements correspond to expression levels (in RPKM units). To mitigate the effect of outliers, we mapped each expression level under the transformation $x := \log(x + 1)$. To draw more direct insights related to GO function prediction, we only considered transcripts whose corresponding genes have known annotations in the GO database. This reduced the number of transcripts from 192,000 to approximately 60,000. Furthermore, we removed transcripts which had uniformly 0 expression levels across all 8,555 samples. After this step, the matrix contained approximately 55,000 transcripts.

2.2 Data Visualization

We performed embeddings of each tissue's samples to visualize the extent to which gene expression varies across tissues. For computational efficiency, we selected the 10,000 remaining transcripts with the highest variance across all samples, and we generated t-SNE

embeddings of the samples. Figure 1 shows two t-SNE embeddings of GTEx samples based on their gene expression profiles.

Figure 1: t-SNE Embedding of GTEx Samples



In both figures, each point represents a single sample from a particular tissue of an individual. The left plot of Figure 1 includes all samples, and the points are colored by tissue. The plot demonstrates that certain human tissues have characteristic gene expression profiles. For example, the points in the large purple cluster towards the right all correspond to samples of skin. Furthermore, the t-SNE embedding demonstrates that for certain tissues, gene expression varies more widely between individuals. For example, the green points spread around the center of the left plot all belong to mammary tissue.

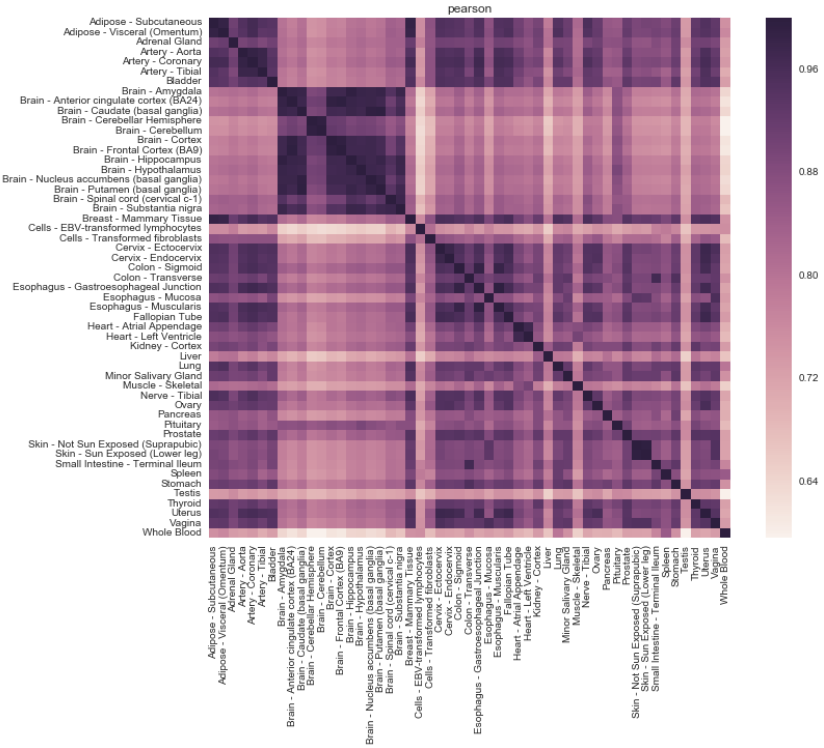
In addition to coloring the samples by tissue, we also generated embeddings that are colored by other information, such as the gender or age of the individual from which the sample was taken. For example, the right plot in Figure 1 shows a t-SNE embedding of gene expression profiles for samples within the brain, colored by age. The heterogeneity of this embedding suggests that age does not significantly impact a person's gene expression profile in the brain.

2.3 Tissue-Tissue Similarity

In order to understand the correlation between tissues in terms of gene expression, we took the median expression level of each transcript in each tissue. This resulted in a vector representing each tissue, where the length of each vector equals the number of transcripts (approximately 55,000). Using these vectors as representations of the 53 tissues, we computed pairwise Pearson correlations and created a tissue-tissue similarity matrix based on gene expression measurements, shown in Figure 2. Tissue Ontology defines tissue similarity based on protein/enzyme functions, while the similarity matrix in Figure 2 is based directly on gene expression. We observed grouping of tissue subtypes in the brain, heart, skin, and artery,

respectively. This indicates that the median expression of a gene within a given tissue may still capture representative characteristics of the tissue.

Figure 2: Tissue-Tissue Correlation by Median Gene Expression



2.4 Obtaining Tissue-Specific GO Functions

We obtained a set of mappings from tissue-specific GO functions to their associated tissues (in Brenda Tissue Ontology format) from a previous work [3]. We then manually curated these mappings to map the GO terms to their associated GTEx tissues, which have a different naming convention than Brenda Tissue Ontology. For each GO function, we then obtained a list of its known gene associations recorded in the GO database. We used the ‘goatools’ python package to obtain the gene sets for the GO queries. We only used GO annotations listed under the following evidence code categories: Experimental Evidence, Computational Analysis, Author Statement, and Curatorial Statement. We did not use Automatically-Assigned annotations, which are less reliable.

After performing the steps described above, we had a list of tissue-specific GO functions. Each prediction task (see *Experiments* below) corresponds to a particular tissue-specific GO function. The number of genes associated with a given GO function dictates the number of examples that may be used in the corresponding prediction task. To avoid having prediction tasks with too few examples, we removed any GO function with fewer than 10 associated genes. This resulted in a list of 266 tissue-specific GO functions, each of which could be used for an independent prediction task.

3. Experiments

3.1 Problem Setup

We addressed whether tissue-specific gene expression can be used to predict associations between genes and tissue-specific biological functions. Thus, for each GO function, we formulated a prediction task in which each gene represents an example whose features are given by the gene's expression levels across the 8,555 GTEx samples. Each example (gene) has a binary label indicating whether or not that gene is associated with the GO function of interest.

For a given GO function, we obtain the positive examples from the list of genes described in Section 2.4. We construct the set of negative gene examples by randomly selecting genes from those genes which are not known to be associated with the biological function of interest. This step assumes that the true underlying gene-function associations are sparse, i.e. most genes and functions are not associated with one another. Note that we select a number of negative examples that is equal to the number of positive examples.

In each prediction task, we split the data with stratified labels into a training set (70%) and a test set (30%). We use the training set to fit the model and tune the hyperparameters, and we report the area under the ROC curve on the test set as our performance metric.

3.2 Feature Reduction

Different tissues contain different numbers of samples in the GTEx dataset. In the prediction task mentioned previously, a varying number of tissue samples translates to different numbers of features for each tissue. This would bias the predictions toward tissues with more samples. To ensure that each tissue has an equal number of features, we performed dimensionality reduction on the set of features within each tissue. In particular, we represented the features for tissue i in a given transcript by the top 5 principal components (PCs) of that transcript in tissue i. (We also used the median expression representation mentioned in Section 2.4 for each problem. As expected, the overall performance using the median expression was not better than when using the top 5 PCs, which can capture variability between individuals.)

3.3 Prediction Tasks

After reducing the dimension of the features, we formulated independent prediction tasks for each GO function. We formulated two different approaches to the prediction tasks: joint-tissue prediction and independent-tissue prediction.

3.3.1 Joint-Tissue Prediction

First, we examined whether gene expression across *all* tissues can be used to predict gene-function associations for tissue-specific GO functions. We applied logistic regression with either the L1 or group lasso penalty. We chose these penalties to ensure that irrelevant features (say, from tissues that turn out to be unrelated) would have 0 coefficients. Using logistic

regression enables us to interpret the relevance of each tissue in a given prediction task by the weights of its coefficients.

To implement logistic regression with group lasso penalty, we used the R package 'grplasso'. We used 5-fold cross validation on the training set to tune the group penalty parameter based on the AUC score. Then we used the penalty parameter with the best AUC score to re-fit the training set and applied this classifier to the test set. The predicted score for each test data point was stored and compared with the true labels using the roc_auc_score from scikit-learn. We followed a similar procedure to run logistic regression with L1 penalty, except that we used scikit-learn throughout.

3.3.2 Independent-Tissue Prediction

Here, we examined whether gene expression in a *single* tissue can be used to predict gene-function associations in various tissues. In this case, each prediction task corresponds to a particular (GO function, tissue) pair. That is, for each GO function, we formulated 53 independent prediction tasks--one for each tissue. In each of these prediction tasks, we used gene expression features from a single tissue to predict which genes are associated with the GO function of interest. Thus, we performed (# GO functions) x (# tissues) = 266 x 53 independent-tissue prediction tasks.

As in the joint-tissue prediction tasks, we used logistic regression with 5-fold cross-validation. However, we used the L2 penalty here. Since the features for a given prediction task all correspond to the same tissue, one should not enforce sparsity of the coefficients.

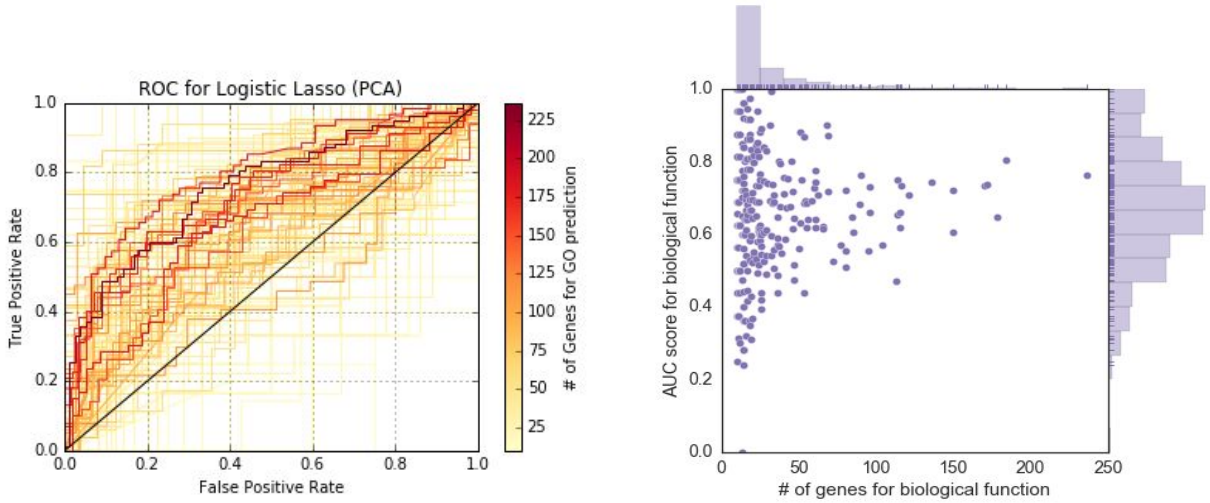
4. Experimental Results

4.1 Joint-Tissue Prediction Results

Figure 3 below shows the results of the joint-tissue prediction tasks using the PCA-reduced expression features (see *Feature Reduction* above). In the left plot, each curve is the ROC curve for a particular GO function's prediction task. The heatmap represents the number of genes associated with a given GO function, which corresponds to the number of positive examples used in the prediction problem. As expected, GO functions with fewer genes (i.e. fewer examples) have more variable ROC curves. The median AUC score across all GO functions for joint tissue-prediction with logistic regression (L1 penalty) was 0.667. The right plot shows these results from a different perspective. Each point corresponds to the AUC score for a given GO function. The histograms show the distributions of gene associations and AUC scores.

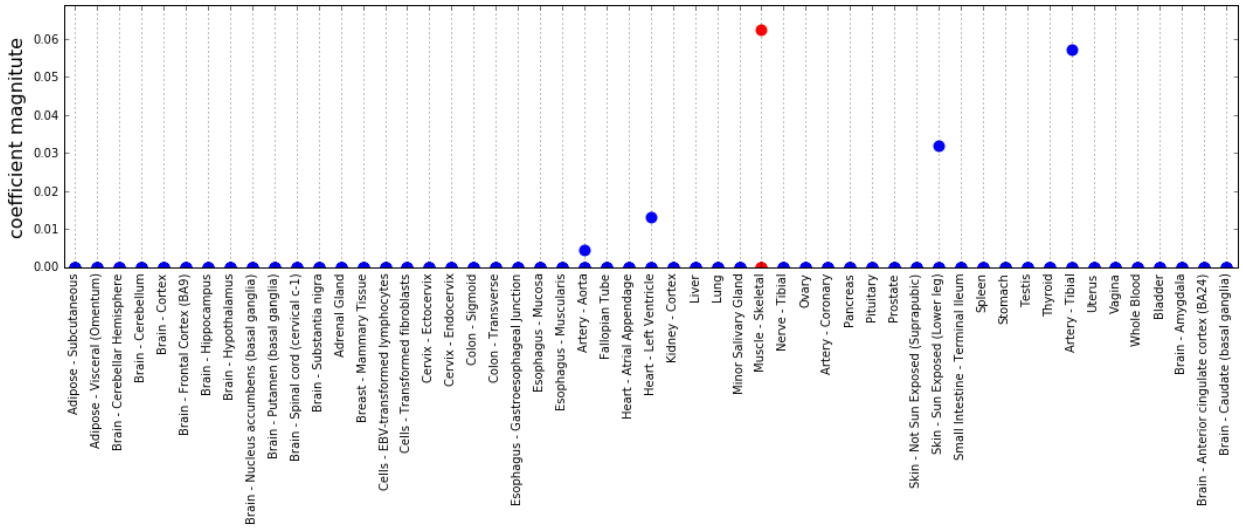
Note that we obtained similar results using logistic regression with the group lasso penalty (median AUC=0.68) and using the median representation of each gene's features (see *Feature Reduction*) for logistic regression with L1 penalty (median AUC=0.65).

Figure 3: Joint-Tissue Prediction using Logistic Regression



We then examined the magnitude of each tissue's feature coefficients. For example, Figure 4 below shows the magnitude of each tissue's coefficients for the prediction task corresponding to the GO function GO:0003009, or "skeletal muscle contraction," which is known from previous work [3] to be associated with the GTEx tissue "Muscle - Skeletal." The column containing the coefficients for skeletal muscle is highlighted in red for emphasis.

Figure 4: Coefficient Weights for GO:0003009 (Skeletal Muscle Contraction), Logistic Regression (L1 Penalty, PCA Features), AUC=0.91



We observe that a feature from skeletal muscle samples attained the highest coefficient weight of any feature. This suggests that gene expression in skeletal muscle is indeed highly informative of certain processes that take place within skeletal muscle. It is also interesting to note that 4 other tissues had a feature with a nonzero coefficient weight: artery (both tibial and aorta), skin (sun-exposed lower leg), and heart (left ventricle).

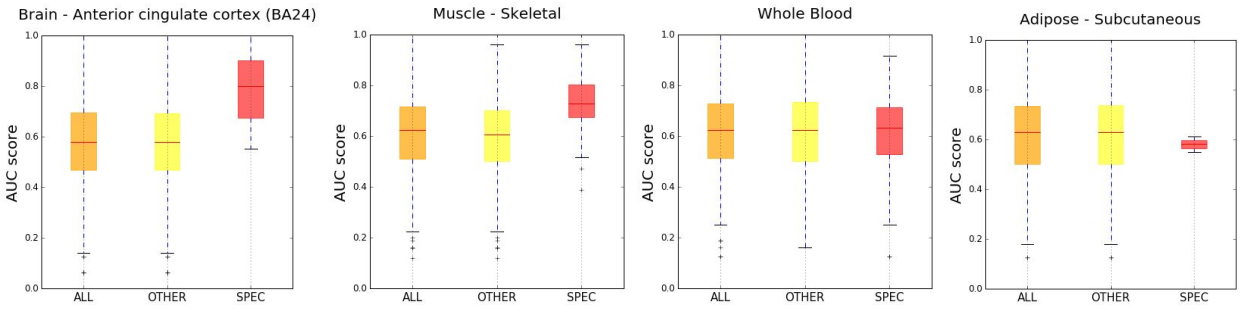
The type of scenario shown in Figure 4 does not always occur. For many GO functions, most tissues have at least 1 feature with a nonzero coefficient weight, making it difficult to determine which tissues were relevant in the prediction task. This probably reflects the fact that some GO functions are more tissue-specific than others.

While the AUC score for joint-tissue prediction (approximately 0.67) is not especially high, it demonstrates that gene expression features can be used to predict gene-function associations. The natural question to address next is whether gene expression features from *individual* tissues can predict gene-function associations.

4.2 Independent-Tissue Prediction Results

Figure 5 below summarizes some of the results for independent-tissue prediction. Each of the 4 boxplots shows the distribution of AUC scores when using features from a particular tissue to predict gene-function associations for either: all 266 GO terms (ALL), GO terms that are not specific to this tissue (OTHER), or GO terms that are specific to this tissue (SPEC).

Figure 5: Independent-Tissue Prediction Results for Several Tissues



We observe that features from “Brain - Anterior cingulate cortex” are significantly better at predicting gene-function associations for GO functions specific to that tissue, when compared to predicting associations for other GO functions. We observe a similar result for “Muscle - Skeletal.” In contrast, when performing an independent-tissue prediction task using features from “Whole Blood,” one does not attain a significantly higher median AUC score for GO functions that are specific to “Whole Blood.” We find a similar result for “Adipose - Subcutaneous,” except that the variance decreases dramatically--this is probably due to the fact that there are only two GO functions associated with “Adipose - Subcutaneous,” while there are 70 GO functions associated with “Whole Blood” in our dataset.

Figure 6 below summarizes the results of the independent-tissue prediction tasks as a “tissue performance” matrix. Each row in the matrix corresponds to the set of GO functions that are specific to a given tissue. Each column corresponds to using gene expression features from a particular tissue. Thus, element (i,j) of the matrix represents the median AUC score when using features from tissue j to predict gene-function associations for GO functions that are known to be specific to tissue i. Note that darker elements correspond to higher AUC scores.

Figure 6: Tissue Performance Matrix

Consider the first row in the tissue performance matrix, which corresponds to GO functions specific to the tibial nerve. We observe that using features from certain individual tissues--even some that are not associated with the nervous system, such as suprapubic skin--results in high AUC scores for predicting gene-function association in the tibial nerve. Now consider the column corresponding to using features from the ovary (10th column from the right). The highest AUC score in this column corresponds to the uterus, suggesting that gene expression in the ovary is informative of biological processes in the uterus.

5. Conclusion and Future Work

We used unsupervised analysis to determine that most tissues have characteristic gene expression profiles. Based on this insight, we then examined which tissues' gene expression profiles are most informative of biological functions in other tissues. We formulated this problem as a prediction task with (dimension-reduced) tissue-specific gene expression levels as features and gene-function associations as responses.

The joint-tissue prediction experiments demonstrated that, for many biological functions, gene expression across all tissues can be used to predict their associations with genes. We used independent-tissue prediction to summarize each tissue's predictive power for each function. This can be used to rank the predictive power of specific tissues for each function.

In the future, we plan to improve the reliability of our prediction tasks by using multiple random negative sets and aggregating the performance. In addition, we could use bootstrap to create a null performance as a baseline to see how likely it is to observe particularly good or bad performance just by chance. For the independent-tissue predictions, we may also use other

supervised algorithms such as support vector machines to independently rank the tissues. We could then test the robustness of our model by comparing this ranking with the ranking obtained from logistic regression with L2 penalty.

In this work, we examined the substitutional effects of using single tissues to predict gene-function associations in other tissues. In the future, we plan to consider complementarity--that is, we plan to use subsets containing multiple tissues in the prediction tasks. Many biological functions involve interactions between multiple tissues, so it would be interesting to see if we can extract such information from our models. These investigations can lead to novel biological interpretations and guide future tissue-specific functional genomics studies.

References

1. Lonsdale, John, et al. "The genotype-tissue expression (GTEx) project." *Nature genetics* 45.6 (2013): 580-585.
2. Ashburner, Michael, et al. "Gene Ontology: tool for the unification of biology." *Nature genetics* 25.1 (2000): 25-29.
3. Pierson, Emma, et al. "Sharing and specificity of co-expression networks across 35 human tissues." *PLoS Comput Biol* 11.5 (2015): e1004220.

Guidelines: http://web.stanford.edu/class/cs341/info.html

The result of the project is a 5-10 page paper. We will **not** accept longer reports.

- Writeup should address :
    - **Problem description:** Give a brief but precise description or definition of the problem or hypothesis you set to evaluate.
    - **Data:** What data did you use?
    - **Experiments:**
        - **Setup:** Describe how did you setup your experiments, how the training/testing data was prepared, what your evaluation methods are, and what baseline methods for comparison are you using.
        - **Results:** Describe your experimental results.
        - **Impediments:** What were some of the issues you encountered in your experiments? How did you deal with them?
    - **Future Work:** What remains to be done?
    - **Brief conclusion**
- Send us the PDF of your final writeup to cs341-spr1516-staff@lists.stanford.edu