Explaining gene expression variation to neural network derived image features

William Jones
Wellcome Trust Sanger Institute
Hinxton
UK, CB10 1SA
wj2@sanger.ac.uk

Abstract

How does gene expression influence visual characteristic of tissues? Can the variation in expression be explaining using data extracting from these images alone? To what extent is this possible? We find that across 10 different tissue types, substation amount of variation of gene expression can be explained using a combination of neural network derived image features and known technical factors. This work indicates that expression biomarkers could soon be well estimated using biomedical images.

1 Introduction

Biomedical images are routinely used by doctors in pathology to diagnose diseases. For example, lung biopsies are taken from patients when grading different levels of lung cancers. Specifically for lung cancer, a great deal of work has focussed on the identification of specific tissue characteristics which are indicative of metastasis. Indeed, this has been the topic of many supervised machine learning competition and approaches. We can now accurately identify metastatic regions from high resolution histopathology lung images, with the state of the art being achieve by Convolutional Neural Networks (CNNs).

What other types of information can be estimated from these high resolution images? To what extent are the visual characteristics of biomedical images influenced by our DNA, and the transcript levels within our tissues? Pathologists are trained to identify visual characteristics in histological slides that are indicative of disease onset. It is known that disease onset triggers changes in bulk RNA expression data in tissues where the diseases present. [1] Furthermore, in many cases that the onset of disease has a genetic component [2].

Owing to the absence such datasets, investigating the interplay between these data modalities intertwined with imaging genetics and histopathology has until been impossible to answer until recently.

We research question by using a dataset of high resolution histopathology images annotated with genotype expression data, and genotypes, as part of the Genotype Tissue Expression Project (GTEx) [3]. At the time of analysis, the repository v6, consists of 449 genotyped individuals, and bulk RNA expression from 44 tissue types [4]. A median of 15 tissues is given per individual, and a median of 155 samples per tissue type. High resolution histopathology images were available for 34 of these tissue types.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

2 Methods Overview

3 Results

We find that across 10 different tissue types, large amount of variation in expression can be explained in large part with a combination of technical factors and image features. Gene expression reflects the biological activity of a cell at a given point in time. Technical factors are already known to have a large impact on gene expression. For example, Ischemic time (SMTISCH), reflects the period time between the donor's death, and tissue extraction. During apoptosis, multiple biological pathways are activated which have a profound impact on the biological activity within a cell. This is reflected by the fact that Ischemic time explains the largest amount of variance in expression out of all of the technical factors. With respect to the image features using the described method, more variance is explained when using large patch-sizes. This is surprising because intuitively one might imagine that variation in gene expression might affect only characteristics in tissues visible at smaller scales. However, these results provide evidence against this intuition.

Futhermore, we compared the effect of using image features extracted from an Inceptionet model without retraining it to differentiate tissue types. We find that we can actually explain more expression variation in this case - perhaps because the output of the raw Inceptionet is a length 2048 vectors, as opposed to a length 1024 vector.

We investigated which technical factors impact both the image feature and expression. RNA degradation number (SMRIN) was the factor that explained the most (11.8%) variation in the image features, wheres Ischemic time explained the most the variance (12.7%) in the expression data. In total, technical factors explained 51.1% of the variation in expression and 40.6% of the variation in the image features.

After regression out the effect of technical covariates from both the image features and expression, we find that for some transcripts as much as 66% of individual transcript can be explained by individual image features. For example (Liver, mean aggregation, retrained at a patch size of 256).

We investigated whether gene transcripts that had a statistically significant amount of variation explained by the image features were enriched for specific biological pathways. For Lung tissue, we find that feature 501 explains a significant amount of variation for gene transcripts that are enriched for the gene ontology term: Defective ABCA3 causes pulmonary surfactant metabolism dysfunction type 3 (SMDP3). This condition is already know to be identifiable by pathologists via histopathology. (Table 1)

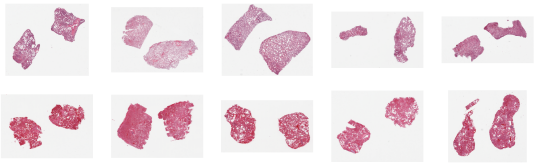
We investigate what visual characteristics individual image features capture. In Lung, a specific image feature that was highly predictive of Ischemic time (the duration of time between donor death and tissue harvesting) appeared to capture the colour (Figure 1b) of the Lung tissue across the tissue area (Figure 1a).

We investigated the genetic basis of variation for individual image features, but found that we did not find any statistically significant associations with genotypes that fall within genes for which the transcript of this gene was significantly associated with expression.

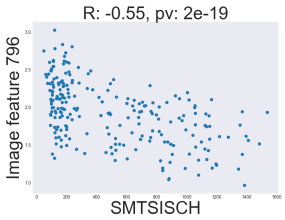
4 Discussion

This study motivates further work that aims to close the information boundary between biomedical images and gene expression. Future work will aim to predict variation in other forms of biomarkers, for example expression of particular genes that might be biomarkers for a given diagnosis. This work will require more comprehensive datasets, but with modern day neural networks able to successfully perform function approximation for a wide variety of complex tasks, this might soon be possible.

5 Figures



(a) Image feature 796 appears to capture global tissue colour.



(b) Image feature 796 strongly predicts Ischemic time.

Figure 1

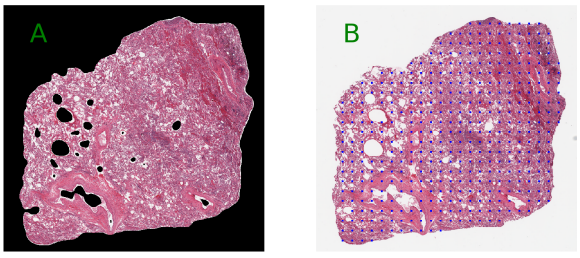


Figure 2: A We illustrate how the tissue is segmented into the tissue foreground and background by Gaussian blurring followed by Otsu thresholding. B We illustrate where patch centers are located within a tissue boundary. The dots represent patch centers that fit inside the tissue boundary.

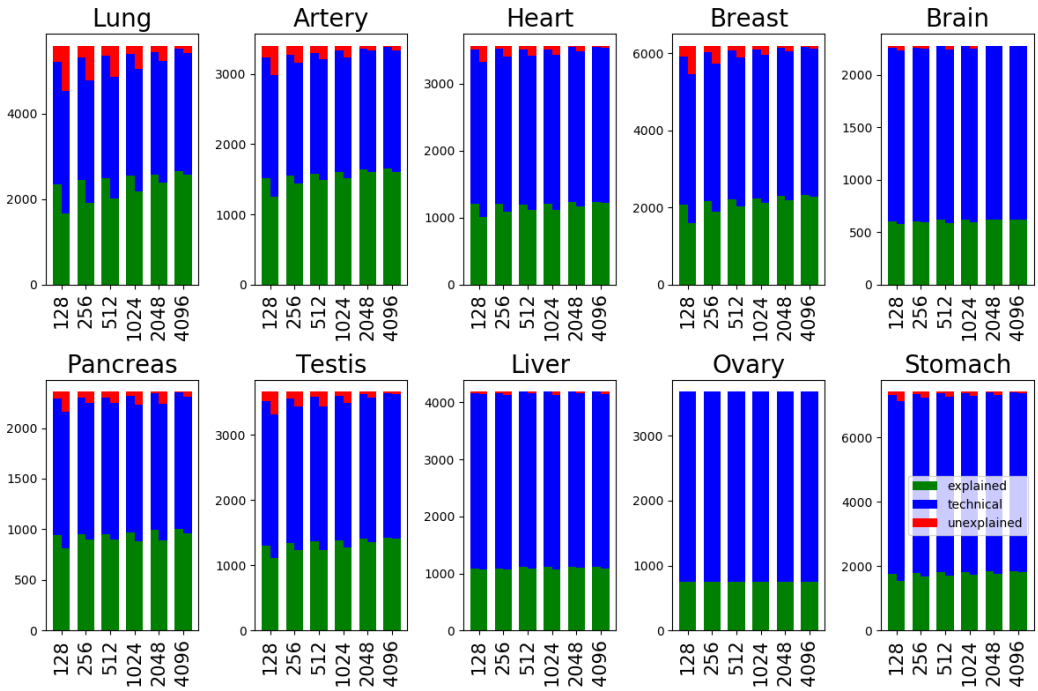


Figure 3

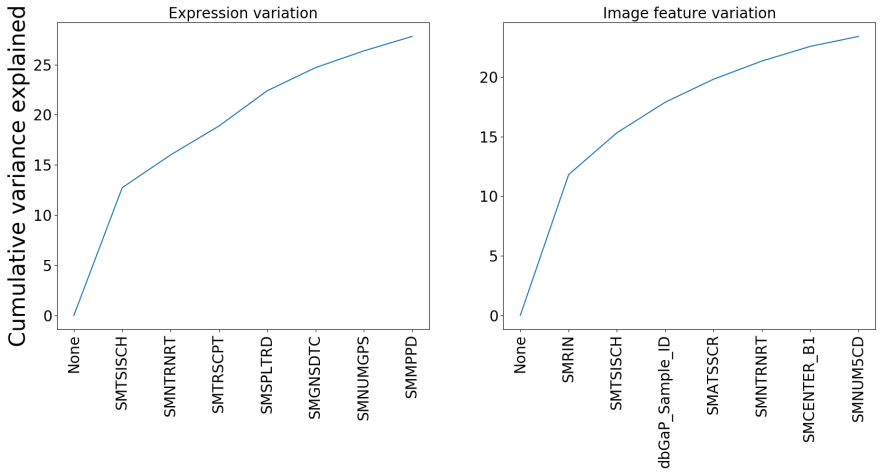


Figure 4: Only displaying the top 8 technical factors. In total, technical factors account for 40.6% of variation in the image features and 51.1% of variation in expression.

Table 1: Enriched Ontology terms for image feature 501

p-value	Ontology	Genes
1.21e-15	lamellar body	LAMP3,CTSH,SFTPA1,NAPSA,SFTPD
6.87e-09	Defective ABCA3 causes pulmonary surfactant metabolism dysfunction type 3 (SMDP3)	SFTPA1,SFTPD,SFTPB,SFTPA2

6 Extended Methods

6.0.1 Training data generation

1. We segment the tissue slice into foreground and background by grayscaling the image, using a Gaussian blur [9] with kernel (51,51), followed by Otsu thresholding [10] (Figure 2).
2. Given a patch size, s, we find all patches of size s that lie within the tissue boundary.

6.0.2 Neural network model

We use Inceptionet-v3, a 220-layer Convolutional Neural Network with pre-trained weights to distinguish everyday objects in images. We follow the common practice of adjusting the network architecture in order to fine-tune the network and repurpose it for a different task [**?**]. To finetune this network, we add a GlobalAveragePooling layer [7] followed by a Dense layer a neural network, and a final softmax layer with 10 classification neurons.

When varying the size of patches, we re-scale all patches to be 299x299 pixels. This means that if we classify a patch of size 4096, the size is drastically reduced to be 299x299. If the patch-size is 128x128, then we use bi-linear interpolation to resize the patch to be 299x299 pixels.

6.0.3 Training and Evaluation

We fine-tuned our modified version of InceptioNet to classify square image patches into their originating tissue types. We use the categorical cross-entropy loss function with Stochastic Gradient Descent with learning rate 0.0001 and momentum = 0.9. We run the back-propagation algorithm to fine-tune the network in the following two steps:

We evaluate the performance of the trained network on a validation set

1. Update the final-layer weights for 10 epochs.
2. Update the InceptionNet-v3 layer weights for 30 epochs.

We assess the performance of the classifier on the held-out validation set by reporting the percentage of correctly classified tissues.

6.1 Latent factors

We choose Lung as the exemplary tissue in which to explore this method. We have a high number of images from this tissue, and the image slices tend to be large. Concretely, we generated image features for each Lung image, using individual lung patches via the following steps.

1. We pass every patch that lies within a tissue boundary through the raw and retrained InceptioNet networks, and at each patch-size, to obtain an image feature vector of length 1024 for each patch.
2. We aggregate the image feature across all image patches using the mean and the median.

In detail, for image i, patch j, I obtain the kth raw patch feature as:

$$r_{ijk} = InceptioNet(x_{ij})_k$$

and, when using the mean aggregation, the final image level features is defined as:

$$f_{ik} = \frac{1}{J} \sum_j r_{ijk}$$

where J is the total number of patches lying within a tissue boundary.

6.1.1 Tools

We used Keras 2.0 [11] to build and train the neural networks. We use OpenSlide Python [12] version 1.1.1 to read in the whole slide images.

6.2 Latent factors

We choose Lung as the exemplary tissue in which to explore this method. We have a high number of images from this tissue, and the image slices tend to be large. Concretely, we generated image features for each Lung image, using individual lung patches via the following steps.

1. We pass every patch that lies within a tissue boundary through the raw and retrained InceptioNet networks, and at each patch-size, to obtain an image feature vector of length 1024 for each patch.
2. We aggregate the image feature across all image patches using the mean and the median.

In detail, for image i, patch j, I obtain the kth raw patch feature as:

$$r_{ijk} = InceptioNet(x_{ij})_k$$

and, when using the mean aggregation, the final image level features is defined as:

$$f_{ik} = \frac{1}{J} \sum_j r_{ijk}$$

where J is the total number of patches lying within a tissue boundary.

6.3 Associating features to RNA

6.3.1 RNA expression data

The RNA expression data was download from the GTEx portal and are recorded in log RPKM values.

6.3.2 Association tests

To investigate strong drivers of variation between the the expression data and the image features, we performed Pearson Correlation tests between the principal components describing the 95% of the variation in the image features and expression respectively. To investigate individual transcript-feature relationships, using the method described in the previous section, we generated sample level features in Lung tissue, for a patch-size of 256x256 pixels with the mean as the aggregation method. We selected the top 500 varying features across all patch sizes, and selected the top 2000 varying transcripts which had mean expression greater than 1. Figure **??** displays where the expression cutoff fall on the histograms of expression mean standard deviations respectively. We investigate the Pearson Correlation tests for each of these pairs or transcript and features (500x2000 in total). These correlations are reported together with a p-value representing the probability that the R score was found by chance.

Plan

Q1: How much variability is shared between image features and expression?

1. ~~Calculate PCA of image features and expression~~
2. ~~Calculate cross correlations between the PCs until 99.9% variance is explained.~~
3. ~~Use the cross correlation measurements to derive total fraction of variance explained of images by expression: calculate $\sum_i \lambda_i \sum_j r_{ij}^2$ where λ_i is the ith eigenvalue of the image feature PCA, r_{ij}^2 is the correlation of image feature PC i with expression feature PC j.~~
4. ~~Make all of the above one-line-to-run for any combination of patch size, feature layer, tissue~~
5. ~~Report the value for each patch size, feature layer, tissue;~~ interpret findings

Q2: How much of the shared variability is due to technical confounders?

1. ~~Regress out known confounders C from the image feature F and expression data E. I.e., fit models $F \sim C + \epsilon$ and $E \sim C + \epsilon$. Use log scale for Ischemic time, monitor all scatter plots for confounders that are included in the model that the linear model assumptions hold. Decide whether to retain full model of all confounders x all genes/features, retain only significant associations, or do forward feature selection.~~
2. ~~Repeat steps Q1:1-5 above using the residuals of this model.~~
3. ~~Create a scatter plot of total shared variance vs. technical shared variance [again, all tissues etc.]~~
4. Interpret findings in context of many tissues, scales, feature layers

Q3: What genes share expression variability to visual features, and at which scales?

1. ~~Using residuals of technical confounders, fit linear model of expression vs. image feature - calculate effect sizes and p-values~~
2. ~~Apply multiple testing correction to derive a list of image feature associated genes.~~
3. ~~Sort genes based on variance explained by each image feature, perform GSEA or gProfiler ranked gene list analysis to test whether particular aspects of biology are enriched.~~ Interpret
4. Assess number of associations [total and per-feature] for different tissues, feature scales. Assess overlap of findings from gene list analyses. Interpret.

Q3A: What genes are influenced by technical confounders?

1. For each confounder, calculate the fraction of variance it explains for each gene expression level.
2. Sort genes based on variance explained by confounder, perform GSEA or gProfiler ranked gene list analysis to test whether particular aspects of biology are enriched.]

Once these done

- Q4: What do the image features represent?
- Q5: Is there a genetic basis to the image features?
- Q6: Are there gene expression levels that cause image feature changes, or vice versa?
- Q7: Are there image features that are associated with clinical annotations? Are any causal?

Results

References

[1] Lewis, P. A., & Cookson, M. R. (2012). Gene expression in the Parkinson's disease brain. Brain Research Bulletin, 88(4), 302?312. http://doi.org/10.1016/j.brainresbull.2011.11.016

[2] Mitchell, K. J. (2012). What is complex about complex disorders? Genome Biology, 13(1), 237. http://doi.org/10.1186/gb-2012-13-1-237

[3] Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., et al. (2013). The Genotype-Tissue Expression (GTEx) project. Nature Genetics, 45(6), 580?585. http://doi.org/10.1038/ng.2653

[4] GTEx Image Viewer https://brd.nci.nih.gov/brd/image-search/searchhome

[5] McCall, M. N., Illei, P. B., & Halushka, M. K. (2016). Complex Sources of Variation in Tissue Expression Data: Analysis of the GTEx Lung Transcriptome. The American Journal of Human Genetics, 99(3), 624ĂŞ635. http://doi.org/10.1016/j.ajhg.2016.07.007

[6] Lee, H., Grosse, R., Ranganath, R., & Ng, A. Y. (2009). Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. Proceedings of the 26th annual pp.609616ĂŞ ACM. http://doi.org/10.1145/1553374.1553453

[7] Lin, M., Chen, Q., & Yan, S. (2013, December 16). Network In Network. arXiv.org.

[8] Baldi, P. (2012). Autoencoders, unsupervised learning, and deep architectures. Presented at the Proceedings of ICML Workshop on Unsupervised and ?.

[9] Shapiro, L. G. & Stockman, G. C: "Computer Vision", page 137, 150. Prentice Hall, 2001

[10] Otsu, N. (1979). A Threshold Selection Method from Gray-Level Histograms. IEEE Transactions on Systems, Man, and Cybernetics, 9(1), 62?66. http://doi.org/10.1109/TSMC.1979.4310076

[11] Keras Chollet, François and others, https://github.com/fchollet/keras

[12] Goode. (2012). OpenSlide: A vendor-neutral software foundation for digital pathology. Journal of Pathology Informatics, 4(1), 27. http://doi.org/10.4103/2153-3539.119005

[13] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative Adversarial Nets. Advances in Neural ?, 2672?2680.

[14] Osokin, A., Chessel, A., Salas, R. E. C., & Vaggi, F. (2017, August 15). GANs for Biological Image Synthesis. arXiv.org.

[15] Kingma, D. P., & Welling, M. (2013, December 20). Auto-Encoding Variational Bayes. arXiv.org.

[16] Long, J., Shelhamer, E., & Darrell, T. (2015). Fully Convolutional Networks for Semantic Segmentation. Proceedings of the IEEE ?, 3431?3440.

[17] Hou, L., Nguyen, V., Samaras, D., Kurc, T. M., Gao, Y., Zhao, T., & Saltz, J. H. (2017, April 3). Sparse Autoencoder for Unsupervised Nucleus Detection and Representation in Histopathology Images. arXiv.org.

[18] Xu, J., Luo, X., Wang, G., Gilmore, H., & Madabhushi, A. (2016). A Deep Convolutional Neural Network for segmenting and classifying epithelial and stromal regions in histopathological images. Neurocomputing, 191, 214?223. http://doi.org/10.1016/j.neucom.2016.01.034

[19] Smith, G. D., & Ebrahim, S. (2008). Mendelian Randomization: Genetic Variants as Instruments for Strengthening Causal Inference in Observational Studies.

[20] Streeter, I., Harrison, P. W., Faulconbridge, A., Flicek, P., Parkinson, H., & Clarke, L. (2017). The human-induced pluripotent stem cell initiative?data resources for cellular genetics. Nucleic Acids Research, 45(D1), D691?D697. http://doi.org/10.1093/nar/gkw928