# Explaining gene expression in histopathology using image features derived from neural networks

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

## Abstract

We investigate to what extent gene expression can be predicted using features extracted from biomedical images. We use data from the Genotype Tissue Expression (GTEx) project comprising high resolution histopathology images with annotated bulk RNA expression and genotype data. We define image features by taking the penultimate layer activations from InceptioNet-v3, after retraining the model to differentiate between 10 tissue types. After aggregating these feature across a tissue segment, we find that in many tissues, and for small patch sizes, we can explain significantly more of the variation in gene expression (>12% more in Lung tissue) than when using permuted image features. We find that technical factors drive a significant proportion of variation in both the image features (40.6%) and gene expression (51.1%). After taking into account technical variation, we find that the variation of certain transcripts can be explained by individual image features (44% for some transcripts in Lung). This work indicates that expression biomarkers could soon be estimated using biomedical images.

## 1 Introduction

Biomedical images are routinely used by doctors in pathology to diagnose disease. For example, Lung biopsies are taken from patients when grading severities of Lung cancers [1]. 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 achieved by Convolutional Neural Networks (CNNs) [2].

Pathologists are trained to identify visual characteristics in histological slides that are indicative of disease, and it is known that disease onset triggers changes in bulk RNA expression data in tissues where the disease presents. [3] Furthermore, in many cases the onset of these has a genetic component [4]. Therefore, it is reasonable to conclude that high resolution images contain information pertaining to the bulk RNA expression profile taken from the sample, and perhaps even the donor genotype.

Until recently the absence such datasets has prohibited work in this direction. With the addition of high resolution histopathology images annotated with gene expression data, and genotypes, as part of the Genotype Tissue Expression Project (GTEx Project) [5], it is now possible to investigate the interplay of histopathology images, gene expression and genetics. At the time of analysis, the repository (v6), consisted of 449 genotyped individuals, along with bulk RNA expression from 44 tissue types [6]. A median of 15 tissues were donated per individual, and a median of 155 samples of each tissue are available in total. 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

### 2.0.1 Training data generation

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

### 2.0.2 Neural network model

We use Inceptionet-v3 [12], 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 [11] 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 reduced to be 299x299. If the patch-size is 128x128, then we use bi-linear interpolation to resize the patch to be 299x299 pixels.

### 2.0.3 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 by first updating the final-layer weights for 10 epoch and then updating the InceptionNet-v3 layer weights for 30 epochs. We used Keras 2.0 [13] to build and train the neural networks. We use OpenSlide Python [14] version 1.1.1 to read in the whole slide images.

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

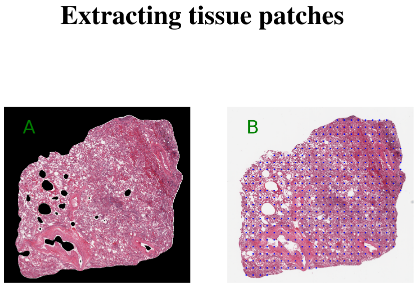

Extracting tissue patches

Figure 1

## 2.1 Aggregated features

We generated image features for each image, using individual lung patches via the following steps: First, 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. Then, we aggregate the image feature across all image patches using the mean.

In detail, for image $i$, patch $j$, I obtain the $k$th raw patch feature as: $r_{ijk} = InceptioNet(x_{ij})_k$

and, after 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.

## 2.2 Associating aggregated features to RNA expression

The RNA expression data was download from the GTEx portal and are recorded in log RPKM values. 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, we generated sample level features in Lung tissue, for a patch-size of 256x256 pixels with the mean as the aggregation method.

### 2.3 Calculating residual features and estimate variance composition

We regress out technical effects from both expression and image features by fitting a linear model using all 51 technical factors as predictors and subtracting the predicted values.

We selected the top 2000 varying transcripts which had mean expression greater than $1$ for each tissue type. We display where the expression cutoff falls on the histograms of expression mean standard deviations respectivel. We investigate the Pearson Correlation tests for each of these pairs or transcript and features. These correlations are reported together with a p-value representing the probability that the R score was found by chance.

We cross correlate the expression and image principle components that explain 99.9% of variation in each respective dataset, and derive the total fraction of expression variance explained by the image features. We calculate $\sum_i \lambda_i \sum_j r_{ij}^2$ where $\lambda_i$ is the $i$th eigenvalue of the image feature PCA and $r_{ij}^2$ is the correlation of image feature PC $i$ with expression feature PC $j$. We recalculate the proportion of variance explained after regressing out known technical factors to derive the proportion of variation explained by the image features, outside of known technical variation.

## 3 Results

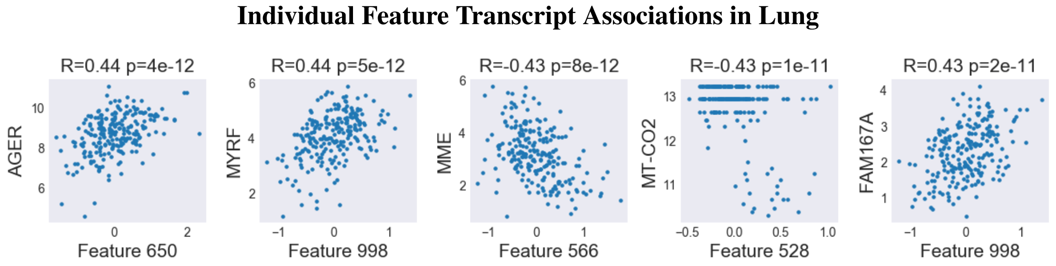


Figure 2

We find that across 10 different tissue types, a large amount of variation in expression can be explained with a combination of known technical factors and image features. (Figure 5). We find that across a number of tissues, more variance is explained using image features from the retrained InceptioNet model than permuted image features.

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 (Figure 3). In total for Lung tissue, technical factors explained 51.1% of the variation in expression and 40.6% of the variation in the image features generated at a patch size of 256.

After regressing out the effect of technical covariates from both the image features and expression, we find that for some transcripts (e.g. AGER) as much as 44% of the variation can be explained by individual image features (Figure 2).

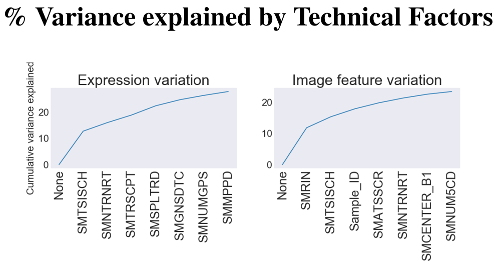
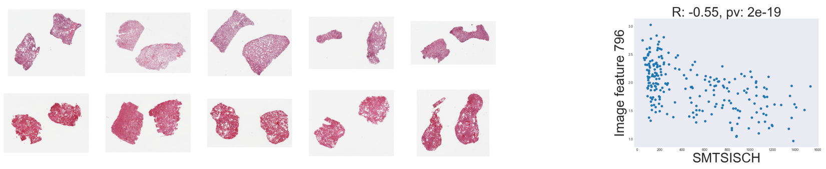


Figure 3

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) (Table 1). This condition is known to be visually identifiable by pathologists via histopathology [7]. Also in Lung tissue, 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 global colour (Figure 4b) across the entire tissue area (Figure 4a).

### Feature Interpretation



(a) Image feature 796 appears to capture global tissue colour. Top row displays top five Lung samples that activate the feature, bottom row displays bottom five.

(b) Image feature 796 strongly predicts Ischemic time (SMTSISCH).

Figure 4

## 4 Discussion

We demonstrate that estimates about gene expression can be estimate using histopathology images features alone. This study motivates further work that aims to close the information boundary between biomedical images and gene expression markers. Future work will aim to predict the variation of specific expression markers for a given disease for 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.

### Variance Components

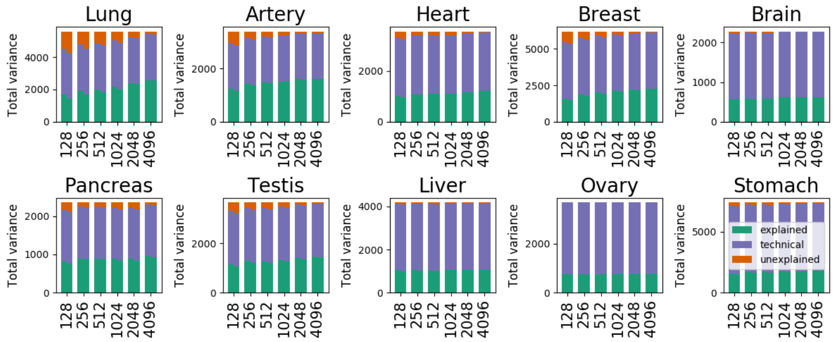


Figure 5: Proportion of the variation in expression explained by image features, explained by technical factors, unexplained, for 10 different tissue types using image features from 6 different patch-sizes (in pixels). For each patch-size, the **left stacked bar** uses retrained InceptioNet features, and **right stacked bar** uses permuted Inception features.

### Enriched Ontology Terms for feature 501

Table 1

| 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 |

## References

[1] Lamb D. Histological classification of lung cancer. Thorax. 1984;39(3):161-165.

[2] Yun Liu & Krishna Gadepalli (2017) Detecting Cancer Metastases on Gigapixel Pathology Images http://arxiv.org/abs/1703.02442v2

[3] 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

[4] 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

[5] 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

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

[7] Susan E. Wert Jeffrey A. Whitsett (2009) Genetic Disorders of Surfactant Dysfunction http://journals.sagepub.com/doi/10.2350/09-01-0586.1

[8] 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

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

[12] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, Zbigniew Wojna (2 Dec 2015). Rethinking the Inception Architecture for Computer Vision https://arxiv.org/abs/1512.00567

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

[14] 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