Explaining gene expression variation using 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 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].

What other types of information and biomarkers can be estimated from these high resolution images? To what extent are the visual characteristics of biomedical images influenced by our DNA, and gene expression levels within our tissues? 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, consists of 449 genotyped individuals, and bulk RNA expression from 44 tissue types [6]. 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

We generated image features using the 1024 final layer activations of an InceptioNet CNN retrained to differentiate between square patches originating from 10 different tissue types. We aggregated these activation across all patches situated within the tissue boundary (Figure **??**). For each image sample, this results in a length 1024 vector representation that we use for downstream association analyses. Detailed methods description contained in the Extended Methods section.

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 known technical factors and image features. (Figure 2)

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.

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, technical factors explained 51.1% of the variation in expression and 40.6% of the variation in the image features.

After regressing out the effect of technical covariates from both the image features and expression, we find that for some transcripts as much as 66% of the variation can be explained by individual image features.

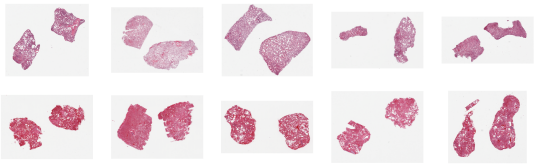
For Lung tissue, we find that feature 501 explains a significant amount of variation for gene transcripts that are enriched for the gene ontology term: Defective ABCA3 causes pulmonary surfactant metabolism dysfunction type 3 (SMDP3) (Table 1). This condition is known to be identifiable by pathologists via histopathology [7]. Also in tissue, a specific image feature that was highly predictive of Ischemic time (the duration of time between donor death and tissue harvesting) appearing to capture the global colour (Figure 1b) across the entire tissue area (Figure 1a).

We investigated the genetic basis of variation for individual image features. However we found that no associations passed Benjamini-Hochberg correction at a 5% FDR rate.

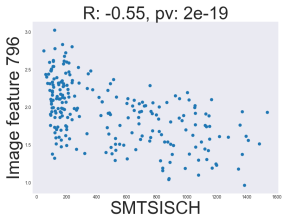
4 Discussion

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.

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 variation of specific markers for a given disease 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.



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



(b) Image feature 796 strongly predicts Ischemic time.

Figure 1

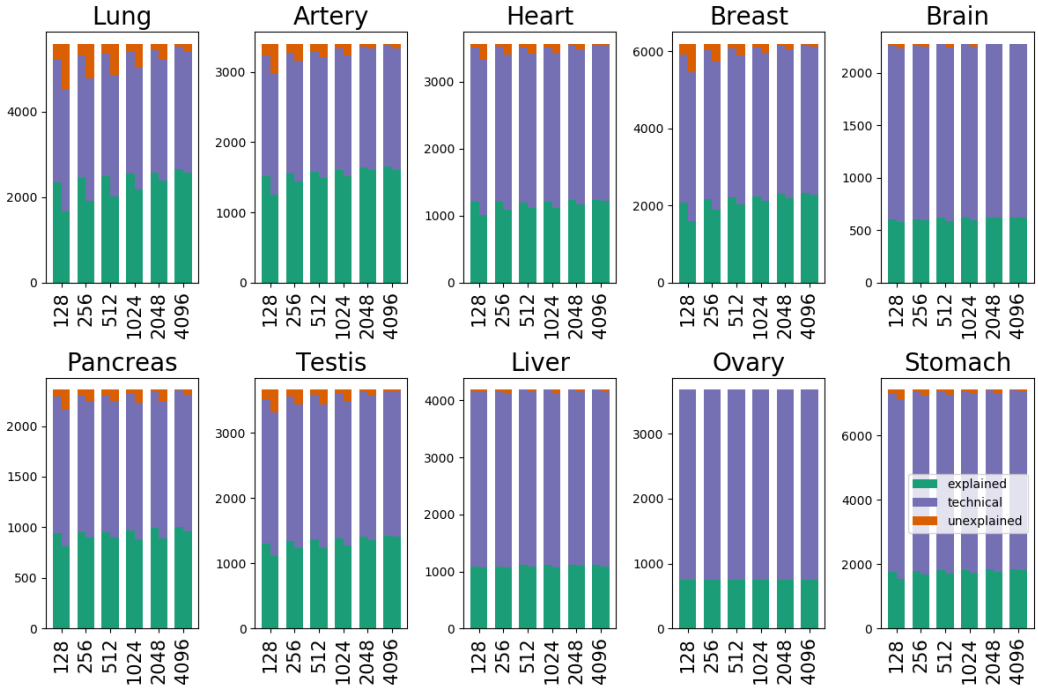


Figure 2: Proportion of the variation in expression explained by image features, explained by technical factors, unexplained, for 10 different tissue types using image features from 6 different patch-sizes. For each patch-size, the left stacked bar uses retrained InceptioNet features, and the right bar uses raw Inception features.

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

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

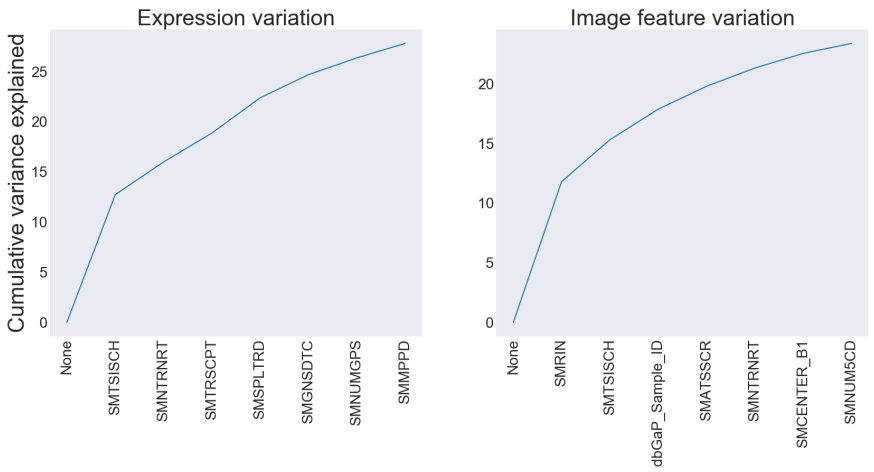


Figure 3: Only displaying the top 8 technical factors. In total, technical factors account for 40.6% of variation in the image features and 51.1% of variation in expression.

[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