Associating gene expression to neural network derived image features

William Jones (First year report)

Abstract 200 words

Finding the genetic basis of biological traits is the central aim of genetics and much of computational biology. For effective and interpretable results, it is important that these traits to be both quantifiable and measurable. Recently, imaging studies have made available large datasets where samples are annotated with extensive genetic and transcriptomic information. Owing to large variation within individual images, and the commonly small sample size, it is difficult to extract well-defined features from images when investigating the genetic background of imaging phenotypes. Existing work has focused on hand-crafted features, however we tackle this challenge with a different approach. We exploit neural networks, known for their ability to extract high-level concepts features from images and use the Genotype Tissue Expression (GTEx) Project high-resolution histology images annotated with bulk gene expression and genotype data. We define a novel scheme to extract global features from an entire histology slide, and compare the number of statistically significant associations found from features at different patch-sizes. We find that much of this variation is due to technical factors, but that there also exists evidence for biologically driven variation.

1 Background 1200 words

The general field to investigate the genetic basis of imaging phenotypes originated through the study of the brain in the 1990s [1] [2] when scientists were searched for the genetic basis of psychopathology. They used using functional neuroimaging to find genes that were active in the brain. Even today, the entire field still broadly maintains this as its focus, indicated by abstract submission for the recent IIGC (International Imaging Genetics Conference), the vast majority of which consider only neuroimaging data and investigate the genetic basis of diseases such as Parkinson's and Alzheimers [3].

In an entirely different field, histopathology, imaging techniques have popularly used in recent years. Histopathology refers to the microscopic examination of tissue in order to study the manifestations of disease. These tissues samples are obtained through surgery, biopsy, or autopsy and then undergo the process of chemical fixation with for-

William Jones
Wellcome Trust Sanger Institute, e-mail: wj2@sanger.ac.uk

malin. After processing and screening, the tissue slices are stained with a combination of hemotoxylin and eosin [4] (often abbreviated H&E). The purpose of the hemotoxylin is to stain nuclei blue, and the purpose of the eosin is the stain the cytoplasm and the extracellular connective tissue matrix pink. This staining gives histopathology images their visual characteristic colours.

Following staining, histological slides are investigated under a microscopy by a pathologist to give a medical diagnosis. Histological slides are used routinely in the clinic to diagnose breast cancer along with numerous other diseases.

Imaging techniques have been used in Computer-assisted diagnosis (CAD) [5] since the 1990s. Since 2010, with the advent of whole slide digital scanners, tissue histopathology slides can now be stored in digital image form. From this point, it became possible to use image analysis and machine learning techniques to complement the opinion of radiologists in disease detection and prognosis prediction.

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. [9] Furthermore, in many cases that the onset of disease has a genetic component [8].

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

The Genotype Tissue Expression (GTEx) project [6] is an example of a dataset that could be used to answer such a problem. This is a repository comprising genetics and bulk RNA expression data from multiple tissues and multiple healthy donors originating from the United States. Early papers studying this data aimed to characterise gene expression variation and quantify eQTLs across different human tissues. At the time of analysis, the repository v6, consists of 449 genotyped individuals, and bulk RNA expression from 44 tissue types. Not every tissue type is available from each donor. A median of 15 tissues is given per individual, and a median of 155 samples per tissue type.

Late last year, in November 2016 - high resolution, histology Whole Side Tissue Images (WSTI) were made available for 34 tissue types [7]. Crucially, each of these samples had annotated donor genotype data, and also the corresponding bulk RNA expression signatures. As such, this resource now had the required level of annotation to study the genetics of histology images.

There have already been recent publications that have already looked at integrating these images with corresponding genetic and transcriptomic data in order to understand pathological phenotypes. McCall et al [10] aim to use image data to reconcile the problem of unknown quantities of cell types within a tissue contributing to the bulk RNA expression of a sample. Within this analysis, they manually annotate the extent of pneumocyte hyperplasia of 114 lung images on a 0-3 scale, and find correlations between these annotations and expression levels of highly variable gene clusters. Although interesting, these image features needed to be handcrafted, a labour intensive and expensive process.

The field of deep learning exploded in notoriety when in 2012 Krizhevsky et al [11] popularized Convolutional Neural networks (CNNs) by emphatically winning the ILSVRC (ImageNet Large Scale Visual Recognition Challenge) with a top-5 test error rate of 15.3% compared to a score of 26.2% achieved by the second best entry. Since this point, they have become the most popular classification method in computer vision research. Part of the reason why they

are so effective, is that they are able to automatically learn feature representations of the classes that are present in the input data. In 2014, Simonyan et al [13] at the Visual Geometry Group in Oxford investigated the final layers of their ILSVRC2013 entry, VGG, and demonstrated that maximizing the activations of neurons deep within the network could generate features similar to the class labels that it was trained upon. The conclusion of this is that discriminatory features from the images of the input data could be learned through the process of training the neural network to classify object classes. From 2014 onwards, a new type of CNN architecture appeared - the Inception architecture [15], based on the ILSVRC2014 winning entry GoogLeNet. This allowed neural networks to become deeper through being efficient with the number of parameters and allowed for more complex representations of imaging data.

Indirectly learning representations via training convolutional networks is not the only way to generate interesting feature representations. Lee et al [12] train convolutional deep belief network on datasets of natural faces, and demonstrate that early layers in the network serve as edge detectors, while middle and late neurons can progressively define complex image features like the contours of eyes, noses, and mouths, with the deepest able to fully the capture an approximate concept of a face.

Images are useful but underutilised source of biological information. Thanks to new techniques, we can now quantify them in useful ways. This leads to questions about the genetic basis of image-based traits, which is underexplored. The GTEX project dataset is the first that allows answering these questions. In this work, we aim to combine the state-of-the-art in each of these fields in order to investigate the interplay between genetics, image analysis and histopathology through the use of cutting edge deep learning techniques.

2 Aims 200 words

In my first year, I aimed to build upon these recent methods from computer vision research in order to relate biomedical images to genetics. Using the GTEx image dataset, I aimed to:

1. Classify tissue type from histology images using neural networks
2. Define a latent image representation using the trained using the trained convolutional neural network as a feature extractor.
3. Understand and interpret these latent factors, their drivers of variation and the relationship of these high level representations to genotype and expression datasets.

3 Methods 500 words

3.1 Tissue Classification

3.1.1 Dataset

We investigated the samples of 10 well-known tissue types with the aim to classify each class based on small square patches extracted from within the tissue boundary. The tissue we considered were: Artery - Tibial, Brain - Cerebellum, Breast - Mammary Tissue, Heart - Left Ventricle, Liver, Lung, Ovary, Pancreas, Stomach and Testis.

3.1.2 Training data generation

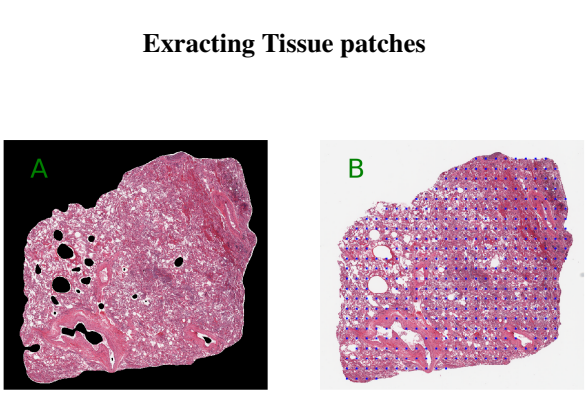


Fig. 1: A We illustrate how the tissue is segmented into the tissue foreground and background by Gaussian blurring followed by Otsu thresholding. B We illustrate where patch centers are located within a tissue boundary. The dots represent patch centers that fit inside the tissue boundary.

Much of the Whole Slide Tissues Image consisted of white empty space, with the area of tissue in question occupying a smaller area in the center. As such, we needed to define a patch sampling strategy that only extracted patches within the tissue boundary. In short, we needed to define the boundary between the tissue foreground and background, with the foreground consisting of the tissue area, and the background consisting of white-space. To do this, we applied the following steps: We assign one-hot encoded tissue labels to the patches, and leave aside one third% of the patches as a validation set.

We chose 6 different pixel widths. 128, 256, 512, 1024, 2048, and 4096. We chose these patch sizes because they represent the diversity of features at different resolutions. Within the patches with width 128 pixels we see individual cell nuclei and clusters of different cell types, whereas in large patch sizes we features characteristic of the entire image, for example colour or tissue shape (Figure 2).

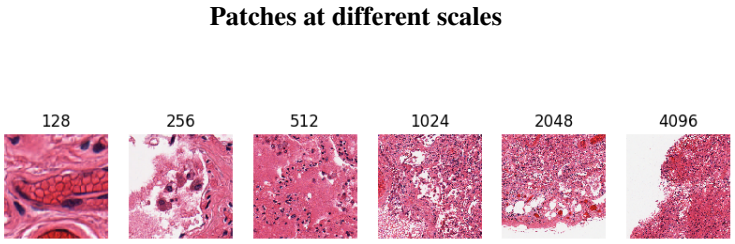


Fig. 2: Resolutions captured by patches of different sizes. For patch sizes at 4096 pixels, whole slide image features are captured such as tissue shape and colour, whereas as smaller resolutions of 128 pixels tissue texture and cell type composition are visible.

1. We segment the tissue slice into foreground and background by grayscaling the image, using a Gaussian blur [17] with kernel (51,51), followed by Otsu thresholding [18] (Figure 1).
2. Given a patch size, *s*, we find all patches of size *s* that lie within the tissue boundary.

We sample 50 image patches within the tissue boundary from 100 images of 10 tissues, giving a total of 5000 patches per image class, with 50000 patches in total.

3.1.3 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 [14]. To finetune this network, we add a GlobalAveragePooling layer [15] 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.

3.1.4 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.

3.1.5 Tools

We used Keras 2.0 [19] to build and train the neural networks. We use OpenSlide Python [20] version 1.1.1 to read in the whole slide images.

3.2 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.

3.3 Associating features to RNA

3.3.1 RNA expression data

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

3.3.2 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 9a 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.

4 Results 1000 words

4.1 Tissue classification

We compared the validation accuracy of using different patch sizes (Figure **??**). We achieved 81% accuracy on held-out test data when using 128x128 patches in training and classification, and 94% when using 4096x4096 image patches, demonstrating that neural networks trained on general tasks can be re-purposed for accurate biological image analysis. Despite the high levels of class ambiguity present at small pixel resolutions, the model is still able to achieve high accuracy. For larger patch sizes, global level features like average pixel intensity and tissue edge boundary seem to give accurate features to determine tissue class.

4.2 Latent Image Feature Association

4.2.1 Exploratory Data analysis

There was also a large amount of heterogeneity in the aggregated image features generated from the histology images across samples (Figure 6a). There exists variation across sample when using both raw and retrained Inceptionet as feature extractors, but that there exists a greater number of active features with retrained Inceptionet. This is characterised by more horizontal red lines. We ask what drives the variation in later sections.

We visualized the features that were generated from each patch to assess which aggregation method would be most appropriate. Figure 5 demonstrates the image feature activations for each patch for two different Lung tissues, with GTEx IDs GTEX-117YW-0526 and GTEX-117YX-1326. Although there exists variation in individual features across patches in a tissue, there are strong drivers that are image specific, and different between each image. Furthermore, we

see that although there is variation across patches for specific features, there exists global levels of activation present in all patches in the image, characteristic of an image level feature, rather than just a patch level features. These are indicated by faint horizontal bands, indicating consistent activation across patches for individual features. As such, we concluded that it would be reasonable to use and compare the median and mean activate across all patches in an image as the aggregation method.

How much redundancy exists in these features? Do they capture independent visual characteristics? We consider image features generated from retrained Inceptionet features, mean aggregated at a patch-size of 256. We cross-correlate these features and perform hierarchical clustering understand their orthogonality and frequency of co-occurrence (Figure 6b). Of the 1024 features, 380 were inactive across all samples. Therefore we excluded these and proceeded using only the remaining 644. We see a large group of co-correlated features, but other groups along the diagonal that are independent.

4.2.2 Feature Associations

To understand whether drivers of variation in expression are related to drivers of variation in the image features, we took the first 10 expression principle components and the first 20 image feature principle components. We performed a pairwise correlation of these major axes of variation (Figure 8). In the top left hand corner, we see that there exists a strong relationship between expression principle component 1 and each of the image feature principle components 1, 3 and 6. These have R score and p-values of $(-0.55, 4^{-23})$, $(0.35, 4^{-9})$, $(0.33, 2^{-8})$ respectively. Furthermore, we see a relationship between image feature PC5 and expression PC2, image feature PC2 and expression PC6. These have R score and p-values of $(0.27, 9^{-6})$, $(-0.25, 3^{-5})$ respectively.

It has been reported that technical factors drive variation in expression [10], and so we investigated the source of this variation within samples. The GTEx dataset records conditions under which data was collected. For example, the total Ischemic time of a sample (SMTSISCH), RNA degradation number (SMRIN) and the code for the center where the sample was collected (SMCENTER). We collected the values of 51 technical features and correlate them with both the image feature PCs and the expression PCs. (Figure 7a and Figure 7b). We see that expression PC1 predominantly captures technical correlation, and is strongly related to the technical factors (Figure 7a). The top 5 ordered correlations between the expression PCs and technical factors all involve PC1. We see that the same technical features correlate with image feature 1 (Figure 7b), albeit in a different order of p-values.

4.2.3 Quantifying individual significant associations

We wanted to investigate whether patch size influenced the core number of associations found between the generated image features and expression. To do this we calculated Pearson Correlation coefficients between the top 500 varying features across all patch sizes, and the top 2000 varying transcripts that had mean expression greater than 1. We filter

both the image features and transcripts because most transcripts do not vary across samples and most image features are inactive across samples.

Strikingly, we observe an interesting observation that across 3 different false discovery thresholds there appears to be an optimum patch-size at 256x256 pixels in order to find Bonferroni significant associations between the aggregated image features and individual transcripts (Figure 9). This is despite the fact that the network was originally retrained to classify size 128x128 pixel patches. The number of associations steadily decreases up to a patch size of 4096, where the number of associations drops to 0.

4.2.4 Comparing Raw vs Retrained InceptioNet

We assessed the degree to which the training process taught the network to detect tissue specific image features (Figure 10). In the same fashion as before, we extracted aggregated image features using raw InceptioNet. This network had been trained to differentiate natural images, but was not exposed to patches from histopathology tissue images. The final layer representation of raw Inceptionet was a vector of length 2048, therefore we filtered two sets of aggregated image features across all lung samples to include only the top 500 varying features across from the retrained and raw network respectively. This was an important step to ensure that we chose the same number of features for both cases. We quantified the number of Bonferroni significant associations across all patches sizes, and at 3 different false discovery rates. We notice that there are significantly more associations found at a patch size of 256 in retrained InceptioNet compared to that of raw InceptioNet. However for larger patch sizes, raw InceptioNet generates a greater number of significant associations, achieving a smaller maximum at patch size 512.

4.2.5 Comparing Aggregation methods

We compare the aggregation methods, mean and median, used to combine features generated from individual patches across the extent of the tissue (Figure 11). We compare the Bonferonni significant associations between image features generated from Lung, at patch-size 256 at three different false discovery rates: 0.01, 0.0001 and 0.000001. Across each FDR the mean aggregate rate generates the greatest number of Bonferroni significant associations. This is true for all patch sizes considered, therefore we concluded that this was the best aggregation method out of the two options.

It might be that single features are strongly associated with the expression of very many genes. This is likely because gene expression is highly correlated amongst groups of genes with a similar biological function. Moreover, the extracted features demonstrate a high degree of correlation (Figure 6b).

4.2.6 Accounting for covarying transcripts and covarying features

We count the number of features with Bonferroni significant associations to at least one transcript (Figure 12). This gives us an indication of how active an image feature is at a particular patch size. We observe a similar behaviour to

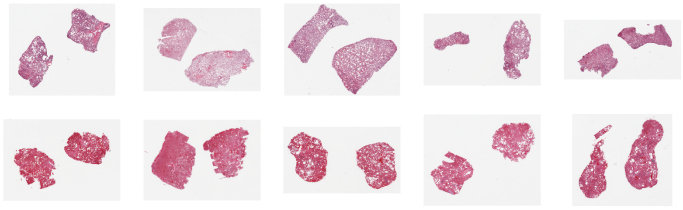
when counting the total numbers of associations. At a patch size of 256 we see the greatest number of features with at least 1 association to a transcript across all FDR thresholds, decreasing consistently for patch sizes greater than 256x256.

In addition, we instead count the number of transcripts with Bonferroni significant associations to at least 1 feature 13. This indicates how many transcripts result in a visual component that is picked up by the generated to image features for each patch size. Unlike before, we observe that the most number of associations is found at patch-size 512, as well as observed a large number of associations at a patch size of 256.

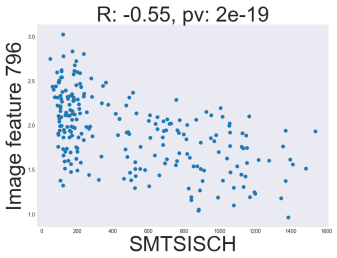
4.2.7 Feature Interpretation

We investigate which image features are associated to these 51 technical factors. We see that image feature 796 is strongly associated to Ischemic time (SMTSISCH) (R -0.55 pvalue 2e-19) (Figure 3b).

We also investigate what this feature represents (Figure 3a) by inspecting the 5 samples that maximally activate and deactivate of feature 796 respectively. We see a clear difference in the colour of the top 5 and bottom 5 samples. This colour change is strongly related to Ischemic time.



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



(b) Image feature 796 strongly predicts Ischemic time.

Fig. 3

5 Discussion 900 words

5.1 Tissue Classification

We have successfully demonstrated that trained convolutional neural networks can be used to classify tissue type from individual square patches. We conclude that accuracy increases with increasing patch-size, suggesting that global level features like average pixel intensity and tissue edge boundary, features that are available at larger scales, seem to give accurate features to determine tissue class.

Although an easy task to a radiologist, it was previously unknown how difficult this task was at smaller resolutions. This work demonstrates that at resolutions as large as 128x128 pixels, there is some ambiguity in predicting tissue

class from image data. This is interesting because we can begin to question which parts of tissues are tissue specific, the proportion of cell types that are shared between tissue types, and how more or less similar certain tissues in the body are.

In this work we consider only 10 tissue types, of very different origin. An extension in this direction would be to include tissue patches from all available tissue classes. Furthermore, the smallest patch-size we investigate is 128x128. It would be interesting to investigate questions such as: the smallest possible patch-size where the validation accuracy is greater than 50%.

5.2 Latent feature associations

Visual characteristics of the images are thought to be driven by levels of bulk RNA expression. We hypothesized that this relationship would be conserved when using the retrained neural network as a feature extractor, and that there would be a quantifiable relationship between these extracted features and transcript levels from bulk RNA expression data. We have shown this to be the case. Broad drivers of variation are shared between the expression datasets and the image features we were generated. Beginning with an Inceptionet neural network containing weights that enable it to classify every day objects, we have demonstrated that we find the greatest number of associations after retraining the neural network to classify tissue type. The maximum is achieved at a patch-size of 256x256, where out of 10 million association tests as using a Bonferroni significance threshold with $\alpha = 0.0001$ we find twice as many significant associations after retraining the network: 31564 compared with 66624 after retraining. Strikingly, for larger patches the trend reverses. For a patch size of 512x512 we find half as many significant associations after retraining: 42235 comparing with 24721 after retraining. The reasons for this observation are likely because the network was retrained to classify tissues type at a patch-size of 128x128, thus focussing the network to be sensitive to patches at these smaller resolutions. This is supported by the observation that the number of significant associations for very large patch-sizes drops sharply after retraining, dropping to 60 significant associations at a patch-size of 4096x4096. One image feature in particular, number 796, has a clear interpretation - it describes the global colour of the tissue. This feature has a strong relationship to Ischemic time and expression PC 1. We conclude from this that the time between death and tissue harvesting not only is a strong driver of variation in RNA expression in the GTEx dataset but that it causes a strong visual effect. This effect should be accounted for and normalised for any future work involving the GTEx image data.

Out of the patch-sizes we have chosen, it is tempting to conclude that the optimum resolution to understand the relationship between RNA expression and image data is at a patch-size of 256. However, after considering just number of transcripts that are significant to a given feature (Figure 13) we see that the story is not so simple. At a patch-size of 256x256 and 512x512, after retraining the network we see that approximately the same number of transcripts are significant to at least one image feature. For 256x256 patches, we find 1960 associations and for 512x512 patches we find 1915. This implies that approximately the same number of transcripts actually have a relationship with the image

features. Since there are approximately 3 times as many associations found after retraining at a patch-size of 256x256 compared to 512x512, (66624 compared with 24721) and that there are approximately the same number of transcripts. This implies that 3 times as many image features are active at a patch-size of 256x256 but result in an equivalent number of interesting associations. We recommend that both patch-sizes should be investigated in future studies of a similar nature, and suggest that patch-sizes smaller than this are too small.

This work demonstrates that it is possible to gain information about gene expression using images alone. We find that variation was largely driven by technical factors, including Ischemic time, RNA degradation, and centre at which the sample was collected.

Many square patch sizes fitted inside the boundary of the tissue slice, and the genetic data available only was representative of the entire image. As such, we needed a method to aggregate the features generated from multiple square patches of the same size, from patches located at different areas of the image into a single feature representative of the entire image. We have only considered two such aggregation methods when many more could have been considered. For example, looking at the maximum activation, or by considering the 90^{th} percentile might be a more appropriate scheme.

Furthermore, using the final layer feature of the neural network compresses entire patches into a single vector of numbers. As such, all spatial structure is lost, and the features are difficult to interpret. We tested methods to investigate the interpretation of these features, such as saliency maps and activation maximisation, but found that neither method returned encouraging results.

6 Future Work 1000 words

6.1 Short term

Given that this report has identified that the strongest drivers of variation between image feature and expression are technical factors, the next step in the analysis is to account for this technical variation in a linear mixed model (LMM) and use this to focus on transcripts with relationships to image features outside of known technical variation. These relationships are more likely to be underpinned biologically and more interesting from a biological perspective. Once these corrected associations are characterised, I will perform Gene Ontology analysis to understand the known biology of these transcripts, and hypothesis as to why they are predictable from the images.

Since we were focussed on correcting for technical factors in order to find and interpret associations independent of these, we have neglected searching for genotype associations. This is because we hypothesised that image features driven by technical factors would drown out the association signal to the genotype, which we already anticipate to be underpowered. A next step would be to investigate the relationship to the genotype of image features that show a relationship to transcripts after correcting for technical factors. I will search for significant associations within a

1kb window of genes which produce transcripts that have significant associations to image features. This will test the hypothesis that image features have a genetic basis.

Once genotype relationships have been identified, I will perform Mendelian Randomization [27] analysis to ascertain whether image features are mediated through expression, or via an independent mechanism.

In addition, I will look to include the clinical annotation data, that were originally made by radiologists when inspecting the tissue slice. I will look for associations with both the generated image features and genotype to investigate whether these clinical phenotype have a genetic basis. Finally, I will extend this analysis to other tissue types. At this point, I have only looked at Lung tissue in detail, but I have pre-computed patches and features from 9 other tissue types, and if interesting finds emerge I can then generate patches from other available tissues.

6.2 Medium term

In the medium term, having performed the corrected p-value analysis, genotype association analysis, integrated clinical phenotypes, performed cross-tissue analysis and performed the mendelian randomisation analysis, I aim to write up the Latent Factor association work into a paper. This will be one of the first example of employing methods from deep learning on images with the purpose of relating them to genetic data. I know of a competing group that is employing similar methods, and so will aim to write this work up as soon as the results become available.

Furthermore, I would like to draft the results of the tissue classification part of this report into a short paper. It would be informative to know at which resolution tissue types become indistinguishable. It is not known whether features at very small resolutions, like colour, can be discriminative. This will answer questions of how different tissues are visually, and could potentially be linked to stages of embryonic development. As an unstudied question this work could answer the resolution at which tissue slices become unique. Furthermore, it flips the conventional use case of neural networks in classification. Instead of testing whether a neural network is able to find discriminative features for a particular task, it assumes that such features would be discoverable by a neural network, if indeed they exist. We aim to investigate whether these feature exist at all, or at least that if they do exist, they cannot be modelling by a neural network.

6.3 Long term

The deep aggregated features that I generated are not easily interpretable. The most interpretable feature only managed to capture the global change of colour in the tissue slice. The reason for this was because the current network architecture collapsed all convolutional layers into a single dense layer, destroying all spatial relationships between the pixels in the input patch. Modern day autoencoders used in computer vision employ Fully Connected Networks (FCNs) [24] architectures for encoding images into compressed representations that are rectangular in shape, and that conserve these spatial relationships. Recently, researchers have used these architectures to perform unsupervised nu-

cleus detection [25] and segmentation of epithelial and stromal regions [26]. Applying these methods to the GTEx data has the potential to give accurate approximations to the proportion of different cell types e.g. lymphocytes in a histopathology tissue slice. Using this, we can whether a relationship exists between cell-type proportion and specific gene transcripts or genotypes. If these features were representative of cell type composition across a tissue slice, it is very likely be related to levels of individual transcripts. This is because the RNA expression data is generated in bulk from a sample taken from the tissue slice, and gene expression is strongly determined by cell-type. It is highly plausible that this would also be linked between to genotype. The benefit of this approach is that these representations actually have a clear interpretation as the proportion of cell-type composition.

I would like to pursue other classes of neural network models in the recent deep learning literature that have given promising results. Generative Adverserial Networks (GANs) [21] and Variational Autoencoders [23] are able to define latent image representations, and creating realistic images with them as input. For example, recent work has demonstrated that it is possible to generate realistic looking images of protein abundance classes in high-content screening [22]. These results demonstrate that it is possible to create more powerful compression mappings images and vectors. The vectors we have looked at in this report are in no way generative, they only describe characteristics of the images. Well trained VAEs and GANs describe the high-dimensional manifold of the compressed image representation that they have the ability to generate entirely new images of a particular class by making small modifications to this compressed representation. This implies that these representations will much better describe the characteristics of the input image when compared with the methods we have currently explored.

6.4 Gantt chart

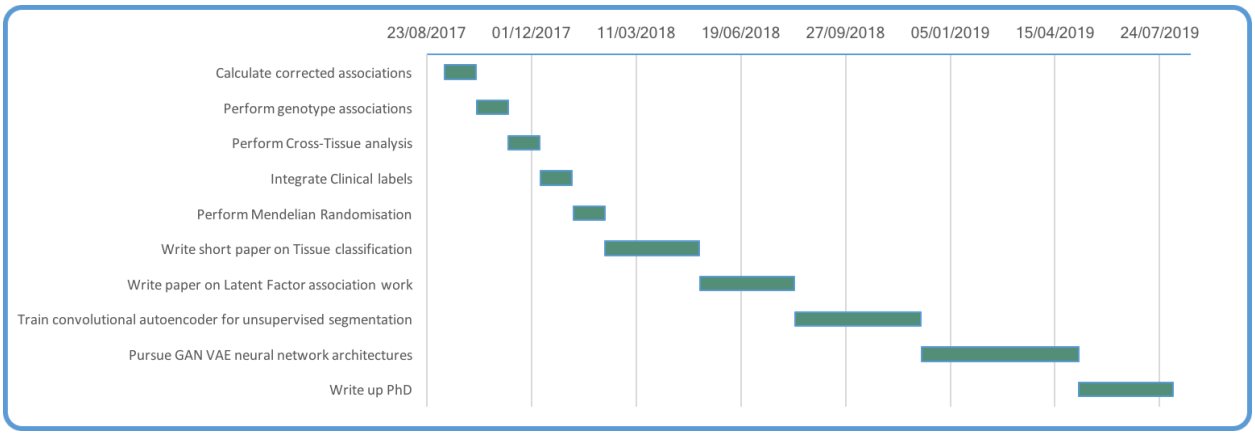


Fig. 4: Gantt chart

7 Figures

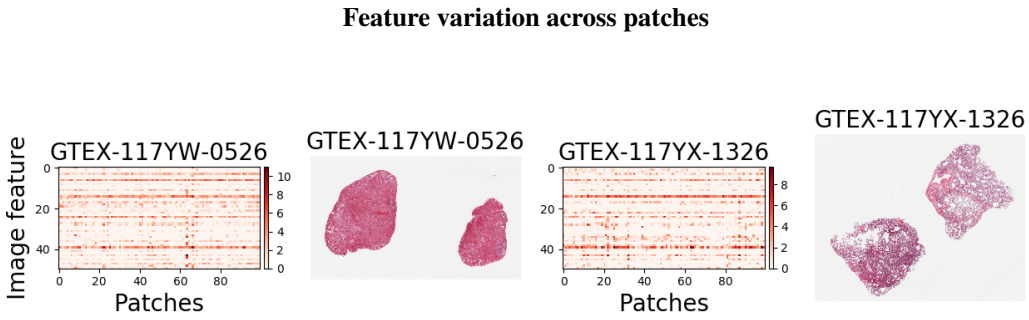
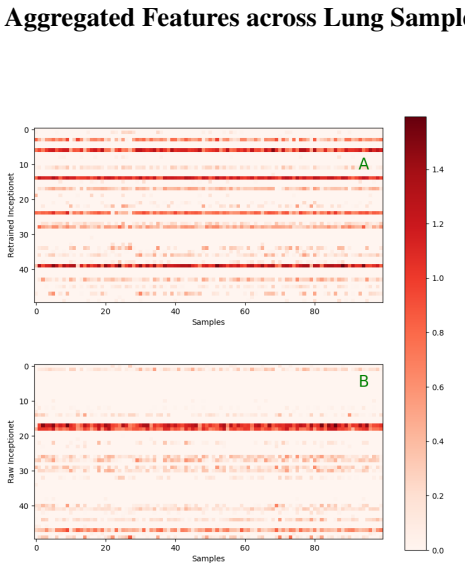
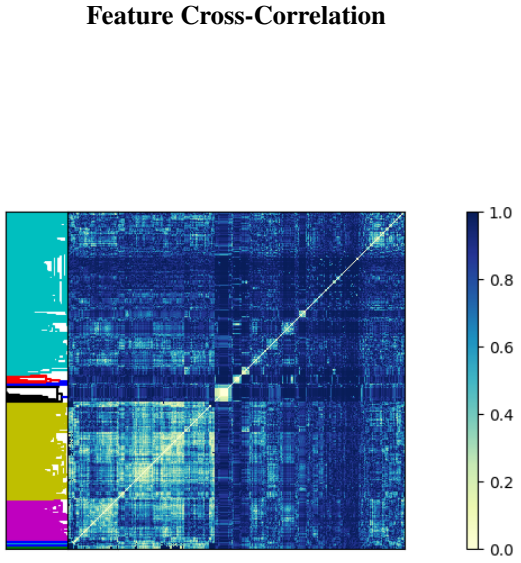


Fig. 5: Variation in 50 aggregated image features across 100 patches from two different Lung samples. Thumbnails of the images are displayed on the right hand side. Despite visible feature variation across patches, there is evidence of image feature activity that is consistent across patches, indicated by the faint horizontal lines. This motivates aggregating across patches, using either the mean or the median.

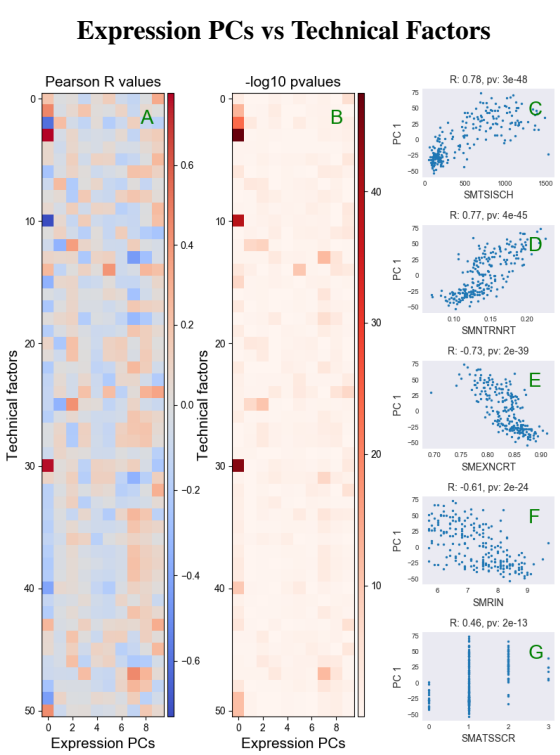


(a) 50 aggregated image feature values across 100 lung image samples. The image features are mean aggregated and generated from the retrained Inceptionet model at a patch size of 256 for retrained Inceptionet A , and raw Inceptionet B. More features activate in retrained Inceptionet.

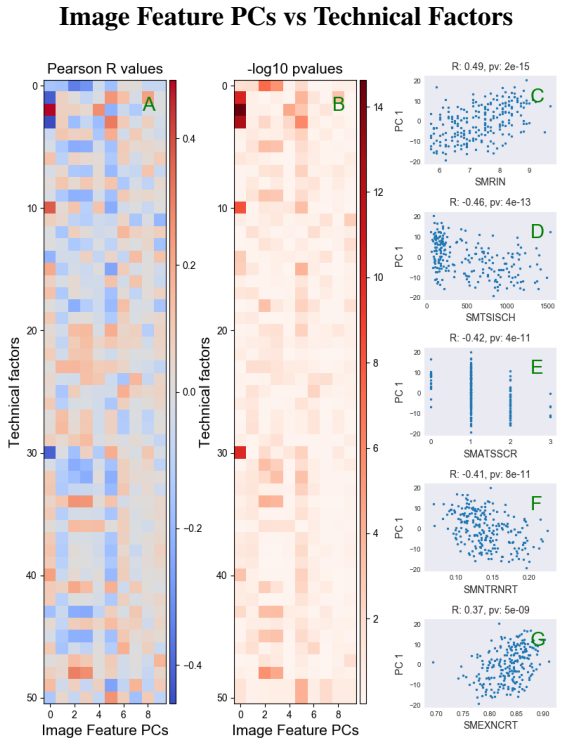


(b) Hierachical clustering performed on aggregated image features with non-zero standard deviation.

Fig. 6



(a) A strong relationship exists between expression PC1 and 5 technical factors. We display A Pearson correlation coefficient and B -log10 pvalues, between 51 technical factors (y-axis) and the first 10 expression PCs (x-axis). Scatterplots of expression PC1 and C Ischemic time, D Intronic mapping rate, E Exonic mapping rate, F RNA degradation, G Autolysis score.



(b) A strong relationship exists between expression PC1 and 5 technical factors. We display A Pearson correlation coefficient and B -log10 pvalues, between 51 technical factors (y-axis) and the first 10 image feature PCs (x-axis). Scatterplots of expression PC1 and CRNA degradation, D Ischemic time, E Autolysis score, F Intronic mapping rate., G Exonic mapping rate.

Fig. 7: Major sources of variation in image and gene expression data are not independent

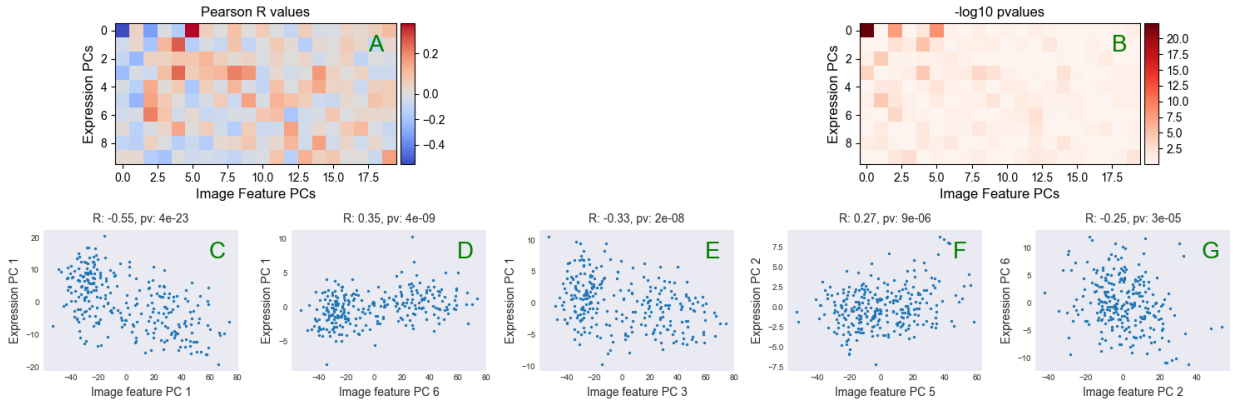
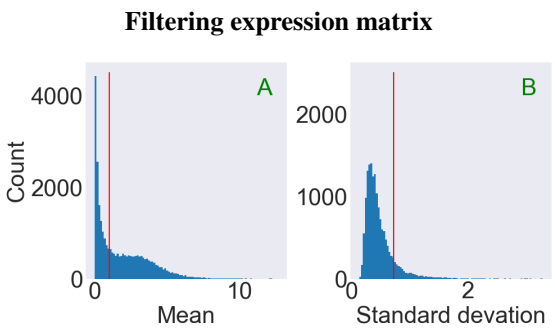
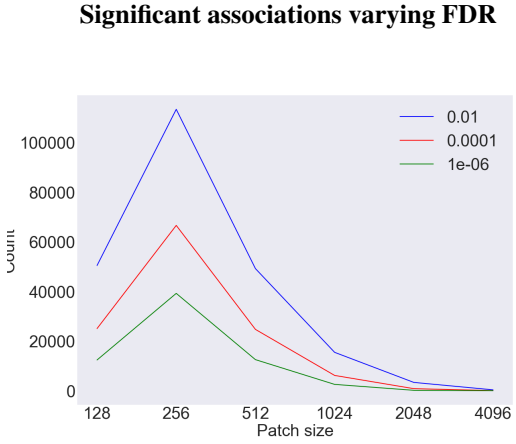


Fig. 8: Major shared variation is between PC1 of both the image features and expression. We display A Pairwise pearsonr correlation and B -log10 pvalues between the top 20 image feature PCs and the top 10 expression PCs. C-G Scatterplot of top 5 (ordered by R^2) associations between expression and image PCs.



(a) Summary statistics of expression levels used for gene filtering A Frequency (y-axis) of mean transcript expression across samples (x-axis). Red vertical line corresponds to x=1, and was used as a minimum expression level cutoff for inclusion in the analysed dataset. B Frequency (y-axis) of the standard deviation of transcript expression across samples (x-axis) that have mean expression greater than 1. Samples to the right of the vertical red line correspond to the 2000 most varying transcripts.



(b) Relationship between raw number of Bonferroni significant associations for 3 different FDRs: 0.01, 0.0001 and 0.000001. We see that at a patch size of 256x256 we find the greatest number of associations across all FDRs.

Fig. 9: The left figure illustrates the expression data cutoffs. The right hand figure displays the number of significant associations between levels of individual transcripts across samples, and the aggregated image features. The image features are from Lung images, mean aggregated, and generated from retrained Inception at a patch size of 256.

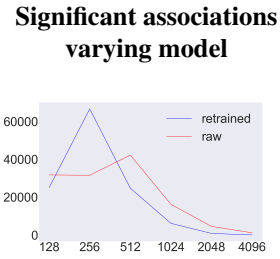


Fig. 10: Comparing the significant association count (y-axis) across patch size (x-axis) between raw and retrained Inceptionet. (Bonf α = 0.0001)

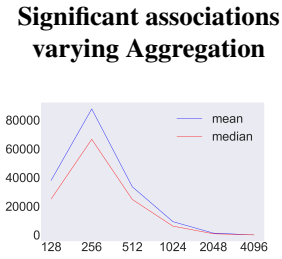


Fig. 11: Comparing the significant association count (y-axis) across patch size (x-axis) between mean and median aggregation. (Bonf α = 0.0001)

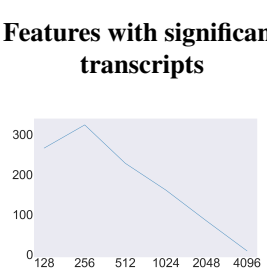


Fig. 12: Counting the number of features that have a significant association to at least 1 transcript. (Bonf α = 0.0001)

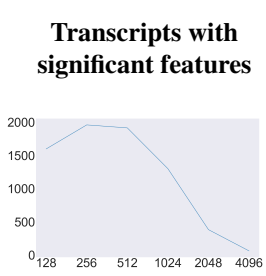


Fig. 13: Counting the number of transcripts that have a significant association to at least 1 feature. (Bonf α = 0.0001)

8 Tables

Table 1: Top 5 R^2 associations between Expression PCs and technical factors

PC	Technical factor	Description	R score	p-value
1	SMTSISCH	Total Ischemic time for a sample in 4 hour intervals	0.78	3e-48
1	SMNTRNRT	Intronic Rate: The fraction of reads that map within introns	0.77	4e-45
1	SMEXNCRT	Exonic Rate: The fraction of reads that map within exons	-0.73	3e-39
1	SMRIN	RIN Number (RNA degradation)	-0.61	2e-24
1	SMATSSCR	Autolysis Score	0.46	3e-13

Table 2: Top 5 R^2 associations between Image Feature PCs and technical factors

PC	Technical factor	Description	R score	p-value
1	SMRIN	RIN Number (RNA degradation)	0.49	2e-15
1	SMTSISCH	Total Ischemic time for a sample in 4 hour intervals,	0.77	4e-13
1	SMATSSCR	Autolysis Score	-0.42	4e-11
1	SMNTRNRT	Intronic Rate: The fraction of reads that map within introns	-0.41	8e-11
1	SMEXNCRT	Exonic Rate: The fraction of reads that map within exons	0.37	5e-09

References

1. Frangou, S., & Murray, R. M. (1996). Imaging as a tool in exploring the neurodevelopment and genetics of schizophrenia. British Medical Bulletin, 52(3), 587?596. http://doi.org/10.1093/oxfordjournals.bmb.a011569
2. MUOZ KE, HYDE LW, HARIRI AR. Imaging Genetics. Journal of the American Academy of Child and Adolescent Psychiatry. 2009;48(4):356-361. doi:10.1097/CHI.0b013e31819aad07.
3. Kim, M., Kim, J., Lee, S.-H., & Park, H. (2017). Imaging genetics approach to Parkinson?s disease and its correlation with clinical score. Scientific Reports, 7, srep46700. http://doi.org/10.1038/srep46700
4. Fischer, A. H., Jacobson, K. A., Rose, J., & Zeller, R. (2008). Hematoxylin and Eosin Staining of Tissue and Cell Sections. Cold Spring Harbor Protocols, 2008(5), pdb.prot4986?pdb.prot4986. http://doi.org/10.1101/pdb.prot4986
5. Gurcan MN, Boucheron L, Can A, Madabhushi A, Rajpoot N, Yener B. Histopathological Image Analysis: A Review. IEEE reviews in biomedical engineering. 2009;2:147-171. doi:10.1109/RBME.2009.2034865.
6. 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
7. GTEx Image Viewer https://brd.nci.nih.gov/brd/image-search/searchhome
8. 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
9. 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

10. 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), 624635. http://doi.org/10.1016/j.ajhg.2016.07.007
11. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural, 10971105.
12. 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. 609616). ACM. http://doi.org/10.1145/1553374.1553453
13. Simonyan, K., Vedaldi, A., & Zisserman, A. (2013, December 20). Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. arXiv.org.
14. Gando, G., Yamada, T., Sato, H., Oyama, S., & Kurihara, M. (2016). Fine-tuning deep convolutional neural networks for distinguishing illustrations from photographs. Expert Systems with Applications, 66, 295301. http://doi.org/10.1016/j.eswa.2016.08.057
15. Lin, M., Chen, Q., & Yan, S. (2013, December 16). Network In Network. arXiv.org.
16. Baldi, P. (2012). Autoencoders, unsupervised learning, and deep architectures. Presented at the Proceedings of ICML Workshop on Unsupervised and ?.
17. Shapiro, L. G. & Stockman, G. C: "Computer Vision", page 137, 150. Prentice Hall, 2001
18. 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
19. Keras Chollet, François and others, https://github.com/fchollet/keras
20. 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
21. 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.
22. Osokin, A., Chessel, A., Salas, R. E. C., & Vaggi, F. (2017, August 15). GANs for Biological Image Synthesis. arXiv.org.
23. Kingma, D. P., & Welling, M. (2013, December 20). Auto-Encoding Variational Bayes. arXiv.org.
24. Long, J., Shelhamer, E., & Darrell, T. (2015). Fully Convolutional Networks for Semantic Segmentation. Proceedings of the IEEE ?, 3431?3440.
25. 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.
26. 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
27. Smith, G. D., & Ebrahim, S. (2008). Mendelian Randomization: Genetic Variants as Instruments for Strengthening Causal Inference in Observational Studies.
28. 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