Identifying Transcriptomic Correlates of Histology using Deep Learning

Liviu Badea[1*], Emil Stănescu[1]

1 Artificial Intelligence and Bioinformatics Group, National Institute for Research and Development in Informatics, Bucharest, Romania

* Corresponding author

Email: badea.liviu@gmail.com

Abstract

Linking phenotypes to specific gene expression profiles is an extremely important problem in biology, which has been approached mainly by correlation methods or, more fundamentally, by studying the effects of gene perturbations. However, genome-wide perturbations involve extensive experimental efforts, which may be prohibitive for certain organisms. On the other hand, the characterization of the various phenotypes frequently requires an expert's subjective interpretation, such as a histopathologist's description of tissue slide images in terms of complex visual features (e.g. 'acinar structures'). In this paper, we use Deep Learning to eliminate the inherent subjective nature of these visual histological features and link them to genomic data, thus establishing a more precisely quantifiable correlation between transcriptomes and phenotypes. Using a dataset of whole slide images with matching gene expression data from 39 normal tissue types, we first developed a Deep Learning *tissue classifier* with an accuracy of 94%. Then we searched for *genes whose expression correlates with features inferred by the classifier* and demonstrate that Deep Learning can automatically derive visual (phenotypical) features that are well correlated with the transcriptome and therefore *biologically interpretable*. As we are particularly concerned with *interpretability* and *explainability* of the inferred histological models, we also develop *visualizations* of the inferred features and compare them with gene expression patterns determined by immunohistochemistry. This can be viewed as a first step toward bridging the gap between the level of genes and the cellular organization of tissues.

Keywords: Deep Learning; digital histopathology; automated visual feature discovery; correlating visual histological features with gene expression data; Deep Learning feature visualization

Abbreviations

H&E hematoxylin and eosin

WSI Whole Slide Image

GTEx Genotype-Tissue Expression project

DL Deep Learning

CNN Convolutional Neural Network

gBP guided Backpropagation

IHC immunohistochemistry

Introduction

Histological images have been used for biological research and clinical practice already since the 19th century and are still employed in standard clinical practice for many diseases, such as various cancer types. On the other hand, the last few decades have witnessed an exponential increase in sophisticated genomic approaches, which allow the dissection of various biological phenomena at unprecedented molecular scales. But although some of these omic approaches have entered clinical practice, their potential has been somewhat tempered by the heterogeneity of individuals at the genomic and molecular scales. For example, omic-based predictors of cancer evolution and treatment response have been developed, but their clinical use is still limited [Xin, 2017] and they have not yet replaced the century old practice of histopathology. So, despite the tremendous recent advances in genomics, histopathological images are still often the basis of the most accurate oncological diagnoses, information about the microscopic structure of tissues being lost in genomic data.

Thus, since histopathology and genomic approaches have their own largely non-overlapping strengths and weaknesses, a combination of the two is expected to lead to an improvement in the state of the art. Of course, superficial combinations could be easily envisioned, for example by constructing separate diagnosis modules based on histology and omics respectively, and then combining their predictions. Or, alternatively, we could regard both histopathological images and genomic data as features to be used jointly by a machine learning module [Mobadersany, 2018].

However, for a more in-depth integration of histology and genomics, a better understanding of the relationship between genes, their expression and histological phenotypes is necessary. Linking phenotypes to specific gene expression profiles is an extremely important problem in biology, as the transcriptomes are assumed to play a causal role in the development of the observed phenotype. The definitive assessment of this causal role requires studying the effects of gene perturbations. However, such genome-wide perturbations involve extensive and complex experimental efforts, which may be prohibitive for certain model organisms.

In this paper we try to determine whether there is a link between the *expression of specific genes* and *specific visual features* apparent in histological images. Instead of observing the effects of gene perturbations, we use representation learning based on deep convolutional neural networks [Goodfellow, 2014] to automatically infer a large set of visual features, which we correlate with gene expression profiles.

While gene expression values are easily quantifiable using current genomic technologies, determining and especially quantifying visual histological features has traditionally relied on subjective evaluations by experienced histopathologists. However, this represents a serious bottleneck, as the number of qualified experts is limited and often there is little consensus between different histologists analyzing the same sample [Baak, 1982]. Therefore, we employ a more objective visual feature extraction method, based on deep convolutional neural networks. Such more objective visual features are much more precisely quantifiable than any subjective features employed by human histopathologists. Therefore, although this correlational approach cannot fully replace perturbational studies of gene-phenotype causation, it is experimentally much easier and at the same time much more precise, due to the well-defined nature of the visual features employed.

In our study, we concentrate on gene expression rather than other multi-omic modalities, since the transcriptional state of a cell frequently seems to be one of the most informative modalities [Yuan, 2014].

There are numerous technical issues and challenges involved in implementing the above-mentioned research.

The advent of *digital* histopathology [Aeffner, 2019] has enabled the large scale application of automated computer vision algorithms for analysing histopathological samples. Traditionally, the interpretation of histopathological images required experienced histopathologists as well as a fairly long analysis time. Replicating this on a computer was only made possible by recent breakthroughs in Deep Learning systems, which have achieved super-human performance in image recognition tasks on natural scenes, as in the ImageNet Large Scale Visual Recognition Competition (ILSVRC) [Krizhevsky, 2012; Russakovsky, 2015]. However, directly transferring these results to digital histopathology is hampered by the much larger dimensions of histological whole slide images (WSI), which must be segmented into smaller image patches (or "tiles") to make them fit in the GPU memory of existing Deep Learning systems. Moreover, since there are far fewer annotated WSI than natural images and since annotations are associated with the whole slide image rather than the individual image tiles, conventional supervised Deep Learning algorithms cannot always be directly applied to WSI tiles. This is particularly important whenever the structures of interest are rare and thus do not occur in all tiles. For example, tumor cells may not be present in all tiles of a cancer tissue sample. Therefore, we concentrate in this paper on normal tissue samples, which are more homogeneous at not too high magnifications and for which the above-mentioned problem is not as acute as in the case of cancer samples.

Only a few large-scale databases of WSI images are currently available, and even fewer with paired genomic data for the same samples. The Cancer Genome Atlas (TCGA) [Weinstein, 2013] has made available a huge database of genomic modifications in a large number of cancer types, together with over 10,000 WSI. The Camelyon16 and 17 challenges [Litjens, 2018] aimed at evaluating new and existing algorithms for automated detection of metastases in whole-slide images of hematoxylin and eosin stained lymph node tissue sections from breast cancer patients. The associated dataset includes 1,000 slides from 200 patients, whereas the Genotype-Tissue Expression (GTEx) [Lonsdale, 2013] contains over 20,000 WSI from various normal tissues.

Due to the paucity of large and well annotated WSI datasets for many tasks of interest (such as cancer prognosis), some research groups employed *transfer learning* by reusing the feature layers of Deep Learning architectures trained on ordinary image datasets, such as ImageNet. Although histopathological images look very different from natural scenes, they share basic features and structures such as edges, curved contours, etc., which may have been well captured by training the network on natural images. Nevertheless, it is to be expected that training networks on genuine WSI will outperform the ones trained on natural scenes [Coudray, 2018].

The most active research area involving digital histopathology image analysis is computer assisted diagnosis, for improving and speeding-up human diagnosis. Since the errors made by Machine Learning systems typically differ from those made by human pathologists [Wang, 2016], classification accuracies may be significantly improved by *assisting* humans with such automated systems. Moreover, reproducibility and speed of diagnosis will be significantly enhanced, allowing more standardized and prompt treatment decisions. Other supervised tasks applied to histopathological images involve [Komura, 2018]: detection and segmentation of regions of interest,

such as automatically determining the tumour regions in WSI [Spanhol, 2016], scoring of immunostaining [Sheikhzadeh, 2016], cancer staging [Wang, 2016], mitosis detection [Shah, 2017], gland segmentation [Chen, 2016], or detection and quantification of vascular invasion [Caie, 2014].

Machine Learning has also been used for discovering new clinical-pathological relationships by correlating histo-morphological features of cancers with their clinical evolution, by enabling analysis of huge amounts of data. Thus, relationships between morphological features and somatic mutations [Coudray, 2018; Chen, 2020; Schaumberg, 2018], as well as correlations with prognosis have been tentatively addressed. For example, [Yu, 2016] developed a prognosis predictor for lung cancer using a set of predefined image features. Still, this research field is only at the beginning, awaiting extensive validation and clinical application.

From a technical point of view, the number of *Deep Learning architectures*, systems and approaches used or usable in digital histopathology is bewildering. Such architectures include supervised systems such as Convolutional Neural Networks (CNN), or Recurrent Neural Networks (RNN), in particular Long Short Term Memory (LSTM) [Hochreiter, 1997] and Gated Recurrent Units [Cho, 2014]. More sophisticated architectures, such as the U-net [Ronneberger, 2015], multi-stream and multi-scale architectures [Kamnitsas, 2017] have also been developed to deal with the specificities of the domain. Among unsupervised architectures we could mention Deep Auto-Encoders (AEs) and Stacked Auto-Encoders (SAEs), Generative Adversarial Networks [Goodfellow, 2014; Gadermayr, 2018], etc.

Currently, CNNs are the most used architectures in medical image analysis. Several different *CNN architectures*, such as AlexNet [Krizhevsky, 2012], VGG [Simonyan, 2014], ResNet [He, 2016], Inception v3 [Szegedy, 2016], etc. have proved popular for medical applications, although there is currently no consensus on which performs best. Many open source Deep Learning *systems* have appeared since 2012. The most used ones are *TensorFlow* (Google [Abadi, 2016]), *Keras* [Chollet, 2015] (high level API working on top of TensorFlow, CNTK, or Theano) and *PyTorch* [Paszke, 2019].

In this paper we look for gene expression-phenotype correlations for normal tissues from the Genotype-Tissue Expression (GTEx) project, the largest publicly available dataset containing paired histology images and gene expression data. The phenotypes consist in visual histological features automatically derived by various convolutional neural network architectures implemented in PyTorch and trained, validated and tested on 1,670 whole slide images at 10x magnification divided into 579,488 tiles. A supervised architecture was chosen, as the inferred visual features should be able to discriminate between the various tissues of interest instead of just capturing the components with the largest variability in the images.

Materials and methods

Our study of the correlations between visual histological features and gene expression involves several stages:

1. Automated *feature discovery* using representation learning based on a supervised Convolutional Neural Network (CNN) trained to classify histological images of various normal tissues; validation and testing of the tissue classifier on completely independent samples.
2. *Quantification* of the inferred histological features in the various layers of the network and their *correlation with paired gene expression data* in an independent dataset.
3. *Visualization of features* found correlated to specific gene expression profiles. Two different visualization methods are used: *guided backpropagation* on specific input images, as well as input-independent generation of *synthetic images* that maximize these features.

The following sections describe these analysis steps in more detail (see also Fig 1).

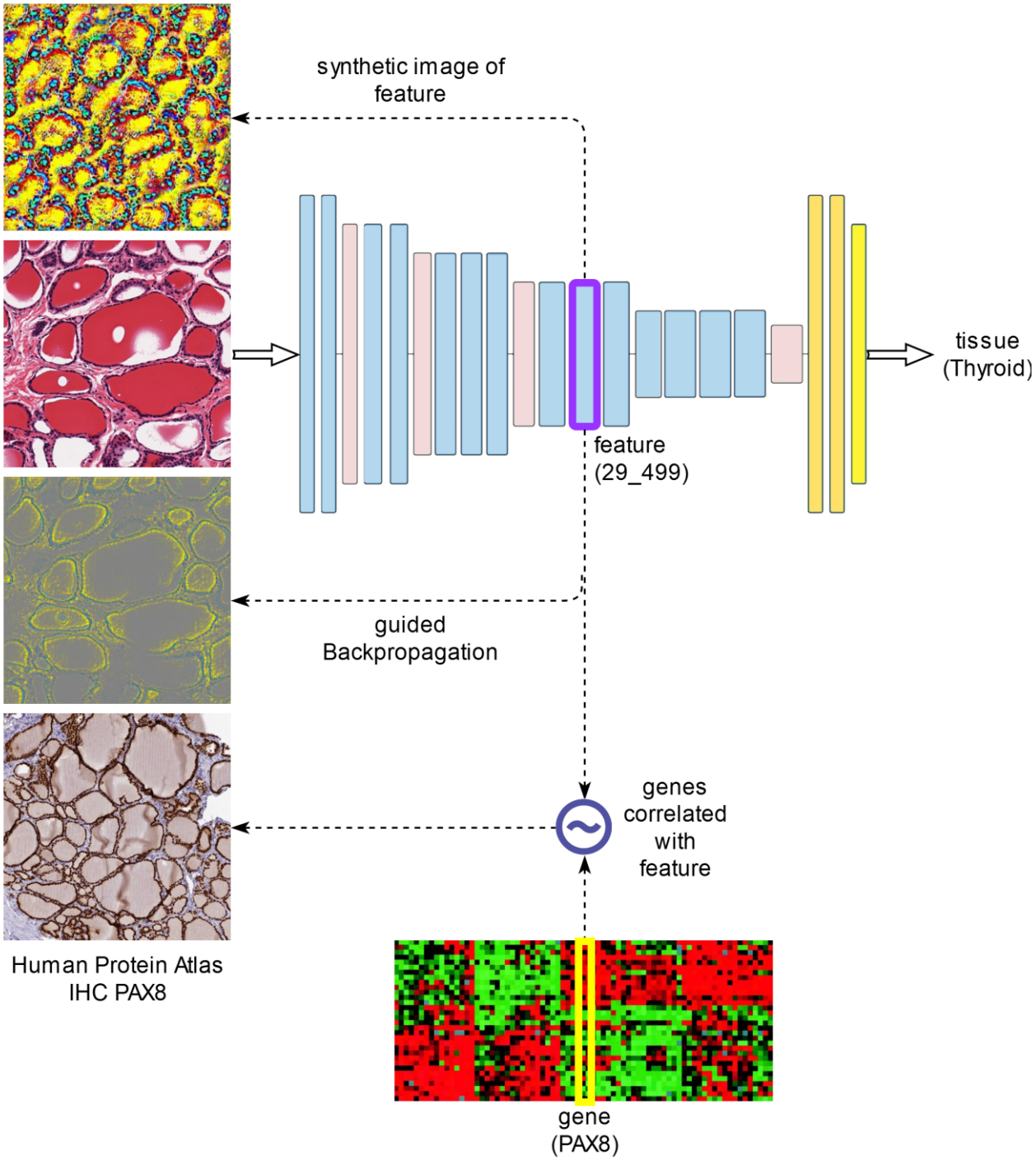


Fig 1. The main analysis steps.

Developing a classifier of histological images

First, we developed a normal tissue classifier based on histological whole slide images stained with hematoxylin and eosin (H&E). Such a tissue classifier is also useful in its own right, since it duplicates the tasks performed by skilled histologists, whose expertise took years or even decades

to perfect. Such a task could not even be reliably performed by computer before the recent advent of Deep Learning. But although Deep Learning has been able to surpass humans in classifying objects from natural images (as in the ImageNet competition [Russakovsky, 2015]), transferring these results to digital histopathology is far from trivial [Komura, 2018]. This is mainly because histopathology images are far larger (up to tens of billions of pixels) and labelled data much sparser than for natural images used in other visual recognition tasks, such as ImageNet. Whole slide images (WSI) thus need to be divided in tiles of sizes small enough (typically 224x224, or 512x512 pixels) to allow Deep Learning mini-batches to fit in the memory of existing GPU boards. Moreover, since there are far fewer annotated WSI than natural images and since annotations are associated with the whole slide image rather than the individual image tiles, conventional supervised Deep Learning algorithms cannot always be directly applied to WSI tiles. This is particularly important whenever the structures of interest are rare and thus do not occur in all tiles. (For example, tumor cells may not be present in all tiles of a cancer tissue sample.) Thus, the main strength of Deep Learning in image recognition, namely the huge amounts of labelled data, is not available in the case of histopathological WSI. More sophisticated algorithms based on multiple instance learning [Xu, 2017] or semi-supervised learning are needed, but haven't been thoroughly investigated in this domain. In multiple instance learning, labels are associated to *bags* of instances (i.e. to the whole slide image viewed as a bag of tiles), rather than individual instances, making the learning problem much more difficult. In contrast, semi-supervised learning employs both labelled and unlabelled data, the latter for obtaining better estimates of the true data distribution.

Therefore, we concentrate in this paper on *normal tissue* samples, which are more homogeneous at not too high magnifications and for which the above-mentioned problem is not as acute as in the case of cancer samples. In the following, we used data from the *Genotype-Tissue Expression* (GTEx) project, the largest publicly available dataset containing paired histology images and gene expression data.

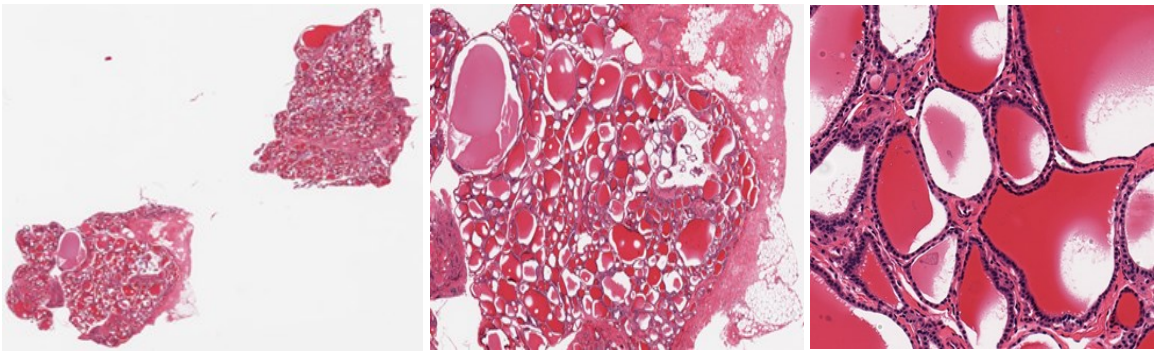
The GTEx dataset

GTEx is a publicly available resource for data on tissue-specific gene expression and regulation [GTEx]. Samples were collected from over 50 normal tissues of nearly 1,000 individuals, for which data on whole genome or whole exome sequencing, gene expression (RNA-Seq), as well as histological images are available.

In our application, we searched for subjects for which both histological images and gene expression data (RNA-Seq) are available. We selected from these 1,670 histological images from 39 normal tissues, with 1,778 associated gene expression samples (for 108 of the histological images there were duplicate gene expression samples). The 1,670 histological images were divided into three data sets:

- *training set* (1,006 images, representing approximately 60% of the images),
- *validation set* (330 images, approximately 20% of images),
- *test set* (334 images, approximately 20% of images).

Figures 2A and 2B show such a histological whole slide image (WSI) and a detail of a thyroid gland sample.



(A) (B) (C)

Fig 2. Whole slide image of thyroid sample GTEX-11NSD-0126.

(A) Whole slide. (B) Detail. (C) Tile of size 512x512.

Another very important aspect specific to WSI is the optimal level of *magnification* to be fed as input to the image classification algorithms. As tissues are composed of discrete cells, critical information regarding cell shape is best captured at high resolutions, whereas more complex structures, involving many cells, are better visible at lower resolutions. The maximum image acquisition resolution in GTEx SVS files corresponds to a magnification of 20x (0.4942 mpp - microns per pixel). In this paper, we chose an intermediate magnification level of 10x (0.9884 mpp), which represents a reasonable compromise between capturing enough cellular details, and allowing a field of view of the image tiles large enough to include complex cellular structures. The whole slide images were divided into tiles of size 512x512 pixels (506μm x 506μm), which were subsequently rescaled to 224x224, the standard input dimension for the majority of convolutional neural network architectures. Image tiles with less than 50% tissue, as well as those with a shape factor other than 1 (with a non-square shape, originating from WSI edges) were removed so as not to distort the training process, and the remaining tiles were labelled with the identifier of the tissue of origin of the whole image. A 512x512 image tile of the WSI from Fig 2A can be seen in Fig 2C.

The image tiles dataset contains 579,488 such tiles (occupying 72GB of disk space) and was further split into distinct training, validation and test datasets as follows:

- 349,340 tiles in the training set (taking up 44GB of disk space),
- 114,123 tiles in the validation set (14GB),
- 116,025 tiles in the test set (15GB).

S1 Table shows the numbers of images and respectively image tiles in the three datasets (training, validation and test) for each of the 39 tissue types. By comparison, the ImageNet dataset (from the ILSVRC-2012 competition) had 1.2 million images and 1,000 classes (compared to 39 classes in the GTEx dataset).

Since to the best of our knowledge, we could not find, for validation purposes, other comparable datasets with paired histological images and gene expression data, we did not perform a color normalization optimized to the GTEx dataset, as it might not extrapolate well to other datasets with different color biases. Therefore, we used the standard color normalization employed in the ImageNet dataset. Such a normalization may be slightly suboptimal, but this is expected to have little impact on the classification results, due to the complexity of the CNNs employed.

Training and validation of Deep Learning tissue classifiers

We trained and validated several different convolutional neural network architectures on the image tile dataset described above (Table 1).

Table 1. Convolutional neural network architectures used in this paper. For ResNet34, the features considered involve only layers that are not "skipped over", while for Inception_v3 only elementary layers with non-negative outputs (ReLU, max/avg-pooling) that are not on the auxiliary branch.

Network	**Number of features** f_{lz}	**Number of trainable parameters**	**Reference**
AlexNet	2,816	57,163,623	Krizhevsky, 2012 [Krizhevsky, 2012]
VGG11	6,976	128,926,119	Simonyan, 2014 [Simonyan, 2014]
VGG13	7,360	129,110,631	Simonyan, 2014 [Simonyan, 2014]
VGG16	9,920	134,420,327	Simonyan, 2014 [Simonyan, 2014] (0)Conv2d(3,64) (1)ReLU (2)Conv2d(64,64) (3)ReLU (4)MaxPool2d (5)Conv2d(64,128) (6)ReLU (7)Conv2d(128,128) (8)ReLU (9)MaxPool2d (10)Conv2d(128,256) (11)ReLU (12)Conv2d (256,256) (13)ReLU (14)Conv2d(256,256) (15)ReLU (16)MaxPool2d (17)Conv2d (256,512) (18)ReLU (19)Conv2d(512,512) (20)ReLU (21)Conv2d(512,512) (22)ReLU (23)MaxPool2d (24)Conv2d(512,512) (25)ReLU (26)Conv2d(512,512) (27)ReLU (28)Conv2d(512,512) (29)ReLU (30)MaxPool2d (31)Linear(25088,4096) (32)ReLU (33)Dropout(0.5) (34)Linear(4096,4096) (35)ReLU (36)Dropout(0.5) (37)Linear(4096,39)
VGG19	12,480	139,730,023	Simonyan, 2014 [Simonyan, 2014]
VGG11_bn	9,728	128,931,623	Simonyan, 2014 [Simonyan, 2014]; Ioffe, 2015 [Ioffe, 2015]
VGG13_bn	10,304	129,116,519	Simonyan, 2014 [Simonyan, 2014]; Ioffe, 2015 [Ioffe, 2015]
VGG16_bn	14,144	134,428,775	Simonyan, 2014 [Simonyan, 2014]; Ioffe, 2015 [Ioffe, 2015]
VGG19_bn	17,984	139,741,031	Simonyan, 2014 [Simonyan, 2014]; Ioffe, 2015 [Ioffe, 2015]
ResNet34	28,992	21,304,679	He, 2016 [He, 2016]
Inception_v3	27,712	24,453,166	Szegedy, 2016 [Szegedy, 2016]
VGG16_1FC	9,920	15,693,159	This paper: same convolutional part as VGG16, but with a single fully connected layer: Conv(VGG16); Dropout(0.5); **Linear(25088,39)**
VGG16_avg1FC	10,432	14,734,695	This paper: the convolutional part as VGG16, followed by an average pooling layer and a single fully connected layer: Conv(VGG16); **AvgPool2d(7,7,512;1,1,512)**; Dropout(0.5); Linear(512,39)

The VGG16_1FC model is a modification of the VGG16 model, in which the final three fully connected layers have been replaced by a single fully connected layer, preceded by a dropout layer. This not only significantly reduces the very large number of parameters of the VGG16 architecture (from 134,420,327 to 15,693,159), but also attempts to obtain more intuitive representations on the final convolutional layers.

Since the last convolutional layer of VGG16 is of the form 7x7 x 512 channels and since location is completely irrelevant in tissue classification, we also extended the convolutional part of VGG16 by

an average pooling layer (over the 7x7 spatial dimensions), followed as in VGG16_1FC by a single fully connected layer. We refer to this new architecture as VGG16_avg1FC.

Architectures with the suffix "bn" use "batch normalization" [Ioffe, 2015].

Note that in the case of ResNet34, the internal layers of the residual blocks do not make up complete representations, since their outputs are *added* to the skip connections. Therefore, the ResNet features considered here involve only layers that are not "skipped over".

On the other hand, the Inception_v3 architecture contains many *redundant* representations, since it combines by concatenation individual multi-scale filter outputs (the individual filter outputs are repeated in the concatenated layer). To avoid this representational redundancy, we only consider as features the *elementary* layers with non-negative outputs (ReLU, max- or average pooling) that are not on the auxiliary branch.

We report, in the following, the accuracies of the classifiers on both *validation* and *test* datasets, after training for 90 epochs on the *training* dataset, with the initial learning rate of 0.01 and reducing it by a factor of 10 every 30 epochs. The mini-batch sizes were chosen according to the available GPU memory (11GB): 256 for AlexNet and 40 for the other models, respectively.

Note that although the validation dataset has not been used for model *construction*, it has been employed for model *selection* (corresponding to the training epoch with the best model accuracy on the validation data). Therefore, we use a completely independent *test* dataset for model evaluation.

The models were implemented in Pytorch 1.1.0 (https://pytorch.org/) [Paszke, 2019], based on the torchvision models library (https://github.com/pytorch/vision/tree/master/torchvision/models).

As the histological images contain multiple tiles, we also constructed a *tissue classifier for whole slide images (WSI)* by applying the tile classifier on all tiles of the given WSI and aggregating the corresponding predictions using the majority vote.

Correlating histological features with gene expression data

From a biological point of view, the visual characteristics of a histological image, its phenotype, is determined by the gene expression profiles of the cells making up the tissue. But the way in which transcriptomes determine phenotypes was practically impossible to determine automatically until the advent of machine learning methods for learning representations based on Deep Learning, mainly because most histological phenotypes were hard to define precisely and especially difficult to quantify automatically.

The internal representations constructed by a convolutional neural network can be viewed as visual features detected by the network in a given input image. These visual features, making up the phenotype, can be correlated to paired gene expression data to determine the most significant gene-phenotype relationships. However, to compute these correlations, we need a precise *quantification* of the visual features.

The *quantification of the inferred visual features* for a given input image is non-trivial, as it needs to be *location invariant* – the identity of a given tissue should not depend on spatial location. Although the activations $Y_{lzxy}(X)$ of neurons (x,y,z) in layer l for input image X obviously depend on the location (x,y), we can construct location invariant aggregates of these values of the form:

$$f_{lz}(X) = \sum_{x,y} Y_{lzxy}(X)^p .$$

We have experimented with various values of p, but in the following we show the simplest version, p=1. This is equivalent to computing the spatial averages of neuron values in a given layer l and channel z. Thus, we employ *spatially invariant features* of the form f_{lz}, which are quantified by forward propagation of histological images X using the formula above.

Since histological whole slide images W were divided into multiple tiles X, we also sum over these tiles to obtain the feature value for W:

$$f_{lz}(W) = \sum_{X \in W} f_{lz}(X) .$$

After quantification of features $f_{lz}(W_i)$ over all samples i, we compute pairwise gene-feature Pearson correlations $r(g,f_{lz})$, where g_i are the gene expression values of gene g in the samples i (for each sample i, we have simultaneous histological whole slide image W_i and gene expression data g_i for virtually all human genes g).

To avoid any potential data leakage, we compute gene-feature correlations on the *test dataset*, which is completely independent from the datasets used for model construction and respectively selection (*training* and respectively *validation* datasets).

Note that since we are in a 'small sample' setting (the number of variables greatly exceeds the number of samples), we performed a univariate analysis rather than a multivariate one, since multivariate regression is typically not recommended for small samples. For example, in the case of the VGG16 architecture, we have 9,920 features f_{lz}, 56,202 genes (totaling 557,523,840 gene-feature pairs) and just 334 samples in the test dataset.

The selection of significantly correlated gene-feature pairs depends on the correlation- and gene expression thresholds used. Genes with low expression in all tissues under consideration should be excluded, but determining an appropriate threshold is non-trivial, since normal gene expression ranges over several orders of magnitude. (For example, certain transcription factors function at low concentrations.)

We compared the different CNN architectures from Table 1 in terms of the numbers of significant gene-feature pairs (as well as unique genes, features and respectively tissues involved in these relations) using fixed correlation- (R_T=0.8) and log_2 expression thresholds (E_T=10). More precisely, the expression threshold E_T constrains the gene expression value E as follows: $log_2(1+E) \leq E_T$.

We also assessed the significance of gene-feature correlations using permutation tests (with N=1,000 permutations).

Since one may expect the final layers of the CNN to be better correlated with the training classes (i.e. the tissue types), we studied the *layer* distribution of features with significant gene correlations.

We also studied the reproducibility of gene-feature correlations w.r.t. the dataset used. More precisely, we determined the Pearson correlation coefficient between the (Fisher transformed) gene-feature correlations computed w.r.t. the validation and respectively test datasets.

Visualization of histological features

The issues of *interpretability* and *explainability* of neural network models is essential in many applications. Deep learning based histological image classifiers can be trained to achieve impressive performance in days, while a human histopathologist needs years, even decades, to achieve similar performance. However, despite their high classification accuracies, neural networks remain opaque about the precise way in which they manage to achieve these classifications. Of course, this is done based on the visual characteristics automatically inferred by the network, but the details of this process cannot be easily explained to a human, who might want to check not just the end result, but also the reasoning behind it. This is especially important in clinical applications, where *explainable* reasoning is critical for the adoption of the technology by clinicians, who need explanations to integrate the system's findings in their global assessment of the patient.

In our context, it would be very useful to obtain a better understanding of the visual features found to be correlated to gene expression, since these automatically inferred visual features make up the phenotype of interest. As opposed to features detected by human experts which are much more subjective and much harder to quantify, these automatically derived features have a precise mathematical definition that allows a precise quantification. Their visualization would also enable a comparison with expert domain knowledge.

In contrast to fully connected network models, convolutional networks tend to be easier to interpret and visualize. We have found two types of visualizations to be particularly useful in our application.

1. The first looks for elements (pixels) of the original image that affect a given feature most. These can be determined using backpropagation of the feature of interest w.r.t. a given input image. In particular, we employ *guided backpropagation* [Springenberg, 2014], which tends to produce better visualizations by zeroing negative gradients during the standard backpropagation process through ReLU units. Such visualizations, which we denote as $gBP(f, X) = \nabla_X^+ f$ depend on the given input image X and were constructed for images from the *test dataset*, which is completely independent from the datasets used for model construction and respectively selection (training and respectively validation datasets).
2. The second visualization is independent of a specific input and amounts to determining a *synthetic input image* that maximizes the given feature $\arg\max_X f(X)$ [Olah, 2017].

Note that unfortunately, guided backpropagation visualization cannot be applied exhaustively because it would involve an unmanageably large number of feature-histological image (f,X) pairs. To deal with this problem, we developed an algorithm for selecting a small number of representative histological image tiles to be visualized with guided backpropagation (see Supporting Information S1 File). The problem is also addressed by using the second visualization method mentioned above, which generates a single synthetic image per feature.

Guided backpropagation visualizations of select features are compared with synthetic images and immunohistochemistry stains for the genes found correlated with the selected feature.

Since the network features were optimized to aid in tissue discrimination, while many genes are also tissue-specific, it may be possible that certain gene-feature correlations are indirect via the tissue variable. To assess the prevalence of such indirect gene(g)-tissue(t)-feature(f) correlations,

we performed conditional independence tests using partial correlations of the form $r(g,f \mid t)$ with a significance threshold p=0.01.

Results

First, we developed a classifier of histological images that is able to discriminate between 39 tissues of interest. A convolutional neural network enables learning visual representations that can be used as visual features in the subsequent stages of our analysis.

The resulting visual features are then quantified on an independent test dataset and correlated with paired gene expression data from the same subjects.

Finally, the features that are highly correlated to genes are visualized for better interpretability.

The following sections describe the results of these analysis steps in more detail.

Deep learning tissue classifiers achieve high accuracies

We trained and validated several different convolutional neural network architectures (from Table 1) on the GTEx image tile dataset. Table 2 shows the accuracies of the classifiers on both *validation* and *test* datasets, after training for 90 epochs on the *training* dataset. Note that the test dataset is completely independent, while the validation dataset has been used for model selection.

Table 2. Classification accuracies for image *tiles*

Network	**accuracy (tiles)**	
	validation	***test***
AlexNet	77.02%	74.76%
VGG11	80.18%	77.49%
VGG13	80.77%	78.19%
VGG16	80.35%	78.00%
VGG19	79.61%	76.93%
VGG11_bn	83.24%	80.74%
VGG13_bn	83.94%	81.56%
VGG16_bn	84.24%	81.46%
VGG19_bn	84.44%	81.54%
ResNet34	82.51%	80.09%
Inception_v3	81.13%	78.99%
VGG16_1FC	82.66%	79.82%
VGG16_avg1FC	83.79%	80.92%

As expected, AlexNet turned out to be the worst classifier among the ones tested. The various VGG architectures produced comparable results, regardless of their size. This may be due to the smaller number of classes (39) compared to ImageNet (1,000). On the other hand, batch normalization leads to improved classification accuracies, but also to representations that show much poorer correlation with gene expression data (see next section).

The modified architectures VGG16_1FC and VGG16_avg1FC perform slightly better than the original VGG16.

Overall, accuracies of 80-82% in identifying the correct tissue based on a single image tile are significant, but still not comparable to human performance. This is because single image tiles allow only a very limited field of view of the tissue.

However, as the images contain multiple tiles, we constructed a *tissue classifier for whole slide images (WSI)* by applying the tile classifier on all tiles of the given WSI and aggregating the corresponding predictions using the majority vote. Table 3 show the accuracies of the whole slide image classifiers obtained using the convolutional neural network models from Table 1.

Table 3. Classification accuracies for *whole slides*

Network	**accuracy (slides)**	
	validation	***test***
AlexNet	91.82%	90.42%
VGG11	93.33%	92.22%
VGG13	93.94%	92.81%
VGG16	92.73%	93.71%
VGG19	93.33%	93.11%
VGG11_bn	93.94%	92.81%
VGG13_bn	93.03%	92.51%
VGG16_bn	93.64%	92.22%
VGG19_bn	93.94%	92.81%
ResNet34	93.33%	92.51%
Inception_v3	92.12%	92.22%
VGG16_1FC	93.33%	93.11%
VGG16_avg1FC	93.64%	94.01%

Note that all architectures, except AlexNet, lead to similar WSI classification accuracies (in the range 92-94%), with the best around 93-94%. (See S4 Table for the associated confusion matrix.) Most of the errors concern confusions between quite similar tissues, such as 'Skin - Sun Exposed' versus 'Skin - Not Sun Exposed'. Merging the following groups of histologically similar tissues (Artery - Aorta, Artery - Coronary, Artery - Tibial), (Colon - Sigmoid, Colon - Transverse), (Esophagus - Gastroesophageal Junction, Esophagus - Muscularis), (Heart - Atrial Appendage, Heart - Left Ventricle), (Skin - Not Sun Exposed, Skin - Sun Exposed) leads to a WSI classification accuracy for 33 tissues of 97.9% in the case of VGG16.

Histological features inferred by VGG architectures correlate well with gene expression data

We calculated the correlations between the visual features of convolutional networks trained on histopathological images and paired gene expression data. We found numerous genes whose expression is correlated with many visual histological features.

Table 4 shows the numbers of significant gene-feature pairs (as well as unique genes, features and respectively tissues involved in these relations) for the various CNN architectures using fixed correlation- (R_T = 0.8) and log_2 expression thresholds E_T = 10 (where $log_2(1+E) \leq E_T$). Histograms of features, genes and their correlations are shown in S1 Fig.

Table 4. Numbers of significantly correlated gene-feature pairs for various network architectures. Fixed correlation- and log_2 gene expression thresholds are used: R_T = 0.8, E_T = 10. Numbers of unique genes, features and tissues involved are also shown (test dataset).

Network	gene-feature pairs	unique genes	unique features	unique tissues
AlexNet	424	55	112	11
VGG11	1,175	70	204	12
VGG13	2,149	70	307	13
VGG16	2,176	74	346	13
VGG19	1,227	71	321	14
VGG11_bn	34	19	7	3
VGG13_bn	59	22	17	6
VGG16_bn	51	20	11	4
ResNet34	3	2	3	2
Inception_v3	23	12	7	2
VGG16_1FC	2,714	69	363	13
VGG16_avg1FC	3,114	83	448	15

Note that although batch normalization slightly improved *tile* classification accuracies (but not *WSI* classification accuracies), it also produced visual representations (features) with dramatically poorer correlation with gene expression data. For example, for VGG16_bn, we could identify only 51 significant gene-feature pairs, involving just 20 unique genes and 11 unique features, compared to 2,176 gene-feature pairs implicating 74 unique genes and 346 unique features for VGG16 without batch normalization.

Unless stated otherwise, we concentrate in the following on visual representations obtained using the VGG16 architecture. (The results for AlexNet and the other VGG architectures without batch normalization are similar.)

VGG16 was selected not just based on it achieving one of the highest WSI classification accuracies on the test dataset (93.71%, Table 3), but also since its last convolutional layer is not directly connected to the output classes (as in the case of VGG16_1FC and VGG16_avg1FC). This is assumed to enable a higher degree of independence of the final convolutional layers from the output classes, rendering them more data oriented.

Table 5 shows the numbers of significantly correlated gene-feature pairs for various correlation- and log_2 gene expression thresholds (for the VGG16 network).

Table 5. Numbers of significantly correlated gene-feature pairs for various correlation- and log_2 gene expression thresholds. Numbers of unique genes, features and tissues involved are also shown (VGG16 network, test dataset).

Correlation threshold	log_2 expression threshold	gene-feature pairs	unique genes	unique features	unique tissues
0.8	10	2,176	74	346	13
0.75	10	4,308	115	647	22
0.7	10	7,984	213	1,146	28
0.75	8	9,624	312	926	26
0.8	7	8,055	365	576	18
0.75	7	15,062	535	1,046	31
0.7	7	27,671	995	1,947	36

Using a correlation threshold $R_T = 0.7$ and a log_2 expression threshold $E_T = 7$, we obtained 27,671 significant correlated gene-feature pairs involving 995 unique genes and 1,947 features (for the VGG16 architecture and the test dataset). We call these gene-feature pairs significant, because the p-values computed using permutation tests (with N=1,000 permutations) were all $p<10^{-3}$.

Fig 3 shows the numbers of significantly correlated genes for the 31 layers of VGG16 (numbered 0 to 30). All features (channels) of a given layer were aggregated, since VGG16 has a too large number of features (9,920 in total). (See also S2 Fig (A) for numbers of correlated genes for *individual* features.) The correlated genes were broken down according to their tissues of maximal expression. Note that the highest numbers of correlated genes correspond to layers with non-negative output (ReLU or MaxPool2d – cf. layer numbers in the VGG16 entry in Table 1).

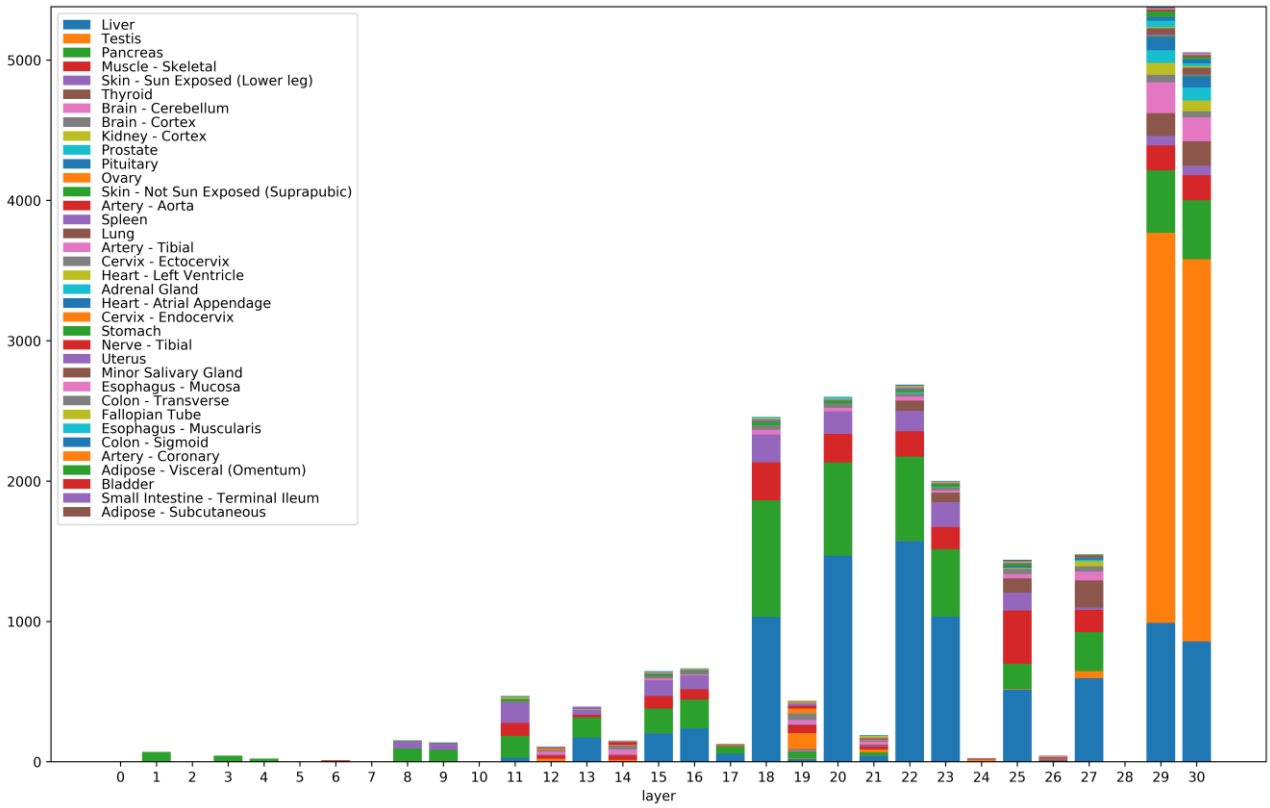


Fig 3. Numbers of significantly correlated genes for the 31 layers of VGG16. Significantly correlated genes were aggregated for all features (channels) belonging to a given layer (i.e. for all features of the form [layer]_[channel]). The correlated genes were broken down according to their tissues of maximal expression. Tissues are color-coded. Colors of the figure bars scanned bottom-up correspond to colors in the legend read top to bottom.

We also remark that the features with significant gene correlations do not necessarily belong to the final layers of the convolutional neural network, even though one may have expected the final layers to be better correlated with the training classes (i.e. the tissue types). Also note that certain tissues with a simpler morphology, such as liver (blue in Fig 3) show correlated genes with features across many layers of the network, from lower levels to the highest levels. Other morphologically more complex tissues, such as testis (orange in Fig 3) express genes correlated almost exclusively with the final layers 29 and 30 (which are able to detect such complex histological patterns). See also examples of histological images for these tissues in Fig 6 (column 1, rows 1 and 2).

For an increased *specificity* and in order to reduce the number of gene-feature pairs subject to human evaluation, we initially performed an analysis with higher thresholds $R_T = 0.8$ and $E_T = 10$ (resulting in 2,176 gene-feature pairs, involving 74 genes and 346 features).

A second analysis with lower thresholds ($R_T = 0.75$, $E_T = 7$) was performed with the aim of improving the *sensitivity* of the initial analysis (15,062 gene-feature pairs, involving 535 genes and 1,046 features - S2 Table).

Since gene-feature *correlations* do not necessarily indicate *causation* (i.e. the genes involved are not necessarily causal factors determining the observed visual feature), we tried to select a subset of

genes, based on known annotations that are more likely to play a causal role in shaping the histological morphology of the tissues. In the following, we report on the subset of genes with Gene Ontology '*developmental process*' or '*transcription regulator activity*' annotations (either direct, or inherited) [Ashburner, 2000; Gene Ontology Consortium, 2019] (Table 6 and S3 Table). While some developmental genes are turned off after completion of the developmental program, many continue to be expressed in adult tissue, to coordinate and maintain its proper structure and function.

Table 6. Developmental and transcription regulation genes correlated with visual features. Genes with Gene Ontology '*developmental process*' or '*transcription regulator activity*' annotations are grouped w.r.t. tissues and ordered by correlation - for each gene we only show the best correlated feature, in the form [layer]_[channel] (for the architecture VGG16 and the test dataset; correlation threshold=0.75, log_2 gene expression threshold=7). Genes with '*transcription regulator activity*' are shown in red.

Gene	Gene name	Tissue with highest expression	Highest median tissue expression (log_2)	Best correlated feature	r
FABP4	Fatty acid-binding protein, adipocyte	Adipose - Visceral (Omentum)	12.6	1_14	0.767
CYP17A1	Steroid 17-alpha-hydroxylase/17,20 lyase	Adrenal Gland	12.4	20_467	0.805
HSD3B2	3 beta-hydroxysteroid dehydrogenase/Delta 5-->4-isomerase type 2	Adrenal Gland	11.2	29_123	0.791
STAR	Steroidogenic acute regulatory protein, mitochondrial	Adrenal Gland	12.5	29_132	0.770
TPM4	Tropomyosin alpha-4 chain	Artery - Aorta	9.1	19_468	0.767
S100A6	Protein S100-A6	Artery - Aorta	11.6	19_162	0.760
YAP1	Transcriptional coactivator YAP1	Artery - Aorta	7.6	19_162	0.758
MARVELD1	MARVEL domain-containing protein 1	Artery - Tibial	7.8	19_162	0.764
GNG12	Guanine nucleotide-binding protein G(I)/G(S)/G(O) subunit gamma-12	Artery - Tibial	7.1	12_209	0.755
ZIC4	Zinc finger protein ZIC 4	Brain - Cerebellum	7.0	29_447	0.922
NEUROD2	Neurogenic differentiation factor 2	Brain - Cerebellum	7.2	29_463	0.915
NEUROD1	Neurogenic differentiation factor 1	Brain - Cerebellum	7.8	29_479	0.873
CRTAM	Cytotoxic and regulatory T-cell molecule	Brain - Cerebellum	7.2	29_479	0.868
ZIC2	Zinc finger protein ZIC 2	Brain - Cerebellum	7.9	29_463	0.854
SLC12A5	Solute carrier family 12 member 5	Brain - Cerebellum	7.5	27_140	0.850
ZIC1	Zinc finger protein ZIC 1	Brain - Cerebellum	8.1	29_479	0.844
ELAVL3	ELAV-like protein 3	Brain - Cerebellum	8.0	29_463	0.834
PVALB	Parvalbumin alpha	Brain - Cerebellum	9.1	29_463	0.781
HPCAL4	Hippocalcin-like protein 4	Brain - Cerebellum	7.5	29_463	0.781
CPLX2	Complexin-2	Brain - Cerebellum	8.7	27_140	0.762
SPOCK2	Testican-2	Brain - Cerebellum	8.6	27_148	0.761
SLC1A2	Excitatory amino acid transporter 2	Brain - Cortex	8.7	18_8	0.844
GRIN1	Glutamate receptor ionotropic, NMDA 1	Brain - Cortex	7.9	18_8	0.812
HPCA	Neuron-specific calcium-binding protein hippocalcin	Brain - Cortex	7.5	29_485	0.811
PACSIN1	Protein kinase C and casein kinase substrate in neurons protein 1	Brain - Cortex	8.5	29_463	0.810
SLC17A7	Vesicular glutamate transporter 1	Brain - Cortex	9.3	18_8	0.799
CAMK2A	Calcium/calmodulin-dependent protein kinase type II subunit alpha	Brain - Cortex	9.2	23_417	0.786
DDN	Dendrin	Brain - Cortex	7.9	25_478	0.775
CEND1	Cell cycle exit and neuronal differentiation protein 1	Brain - Cortex	8.6	29_485	0.755
KIF5A	Kinesin heavy chain isoform 5A	Brain - Cortex	10.0	29_485	0.754
NBL1	Neuroblastoma suppressor of tumorigenicity 1	Cervix - Ectocervix	9.6	19_162	0.758
CRTAP	Cartilage-associated protein	Cervix - Ectocervix	8.1	12_223	0.756
HMGB1	High mobility group protein B1	Cervix - Ectocervix	7.4	19_468	0.753
PIAS3	E3 SUMO-protein ligase PIAS3	Cervix - Endocervix	7.0	19_468	0.764
SPIN1	Spindlin-1	Cervix - Endocervix	7.2	19_333	0.752
PLXNB2	Plexin-B2	Cervix - Endocervix	8.0	12_164	0.750

Gene	Gene name	Tissue with highest expression	Highest median tissue expression (log_2)	Best correlated feature	r
NKX2-5	Homeobox protein Nkx-2.5	Heart - Atrial Appendage	7.0	30_214	0.903
BMP10	Bone morphogenetic protein 10	Heart - Atrial Appendage	8.3	29_313	0.855
NPPA	Natriuretic peptides A	Heart - Atrial Appendage	14.9	29_481	0.777
NMRK2	Nicotinamide riboside kinase 2	Heart - Atrial Appendage	9.8	30_72	0.751
MYBPC3	Myosin-binding protein C, cardiac-type	Heart - Left Ventricle	10.8	23_270	0.783
TNNI3	Troponin I, cardiac muscle	Heart - Left Ventricle	12.0	25_28	0.780
TNNT2	Troponin T, cardiac muscle	Heart - Left Ventricle	11.6	25_31	0.759
CSRP3	Cysteine and glycine-rich protein 3	Heart - Left Ventricle	9.4	30_72	0.755
AQP2	Aquaporin-2	Kidney - Cortex	7.6	29_280	0.890
UMOD	Uromodulin	Kidney - Cortex	7.4	27_274	0.874
PLG	Plasminogen	Liver	8.7	29_234	0.959
APOA5	Apolipoprotein A-V	Liver	8.8	29_192	0.954
SERPINC1	Antithrombin-III	Liver	10.1	29_192	0.952
HRG	Histidine-rich glycoprotein	Liver	9.2	29_234	0.952
F2	Prothrombin	Liver	9.4	29_192	0.949
AHSG	Alpha-2-HS-glycoprotein	Liver	10.7	29_192	0.946
APCS	Serum amyloid P-component	Liver	10.7	29_192	0.942
BAAT	Bile acid-CoA:amino acid N-acyltransferase	Liver	7.7	29_192	0.938
ANGPTL3	Angiopoietin-related protein 3	Liver	7.3	29_234	0.934
APOA2	Apolipoprotein A-II	Liver	12.2	29_192	0.931
CYP4A11	Cytochrome P450 4A11	Liver	8.3	29_234	0.924
CPB2	Carboxypeptidase B2	Liver	8.5	29_192	0.922
PROC	Vitamin K-dependent protein C	Liver	7.9	29_234	0.919
APOH	Beta-2-glycoprotein 1	Liver	11.7	29_234	0.919
FGB	Fibrinogen beta chain	Liver	12.6	29_192	0.896
FGL1	Fibrinogen-like protein 1	Liver	10.2	18_231	0.879
G6PC	Glucose-6-phosphatase	Liver	7.1	29_234	0.872
FGA	Fibrinogen alpha chain	Liver	12.2	30_192	0.858
ASGR2	Asialoglycoprotein receptor 2	Liver	8.8	30_192	0.854
CRP	C-reactive protein	Liver	12.6	29_192	0.852
VTN	Vitronectin	Liver	11.2	18_279	0.851
FGG	Fibrinogen gamma chain	Liver	12.2	30_192	0.829
IGFBP1	Insulin-like growth factor-binding protein 1	Liver	7.2	29_75	0.802
APOB	Apolipoprotein B-100	Liver	8.5	30_438	0.784
CREB3L3	Cyclic AMP-responsive element-binding protein 3-like protein 3	Liver	8.2	20_481	0.777
CPS1	Carbamoyl-phosphate synthase [ammonia], mitochondrial	Liver	8.6	29_192	0.775
ARG1	Arginase-1	Liver	9.0	20_194	0.768
SFTPB	Pulmonary surfactant-associated protein B	Lung	12.2	29_383	0.789
SCGB1A1	Uteroglobin	Lung	9.2	29_155	0.776
STATH	Statherin	Minor Salivary Gland	7.7	25_200	0.767
MYF6	Myogenic factor 6	Muscle - Skeletal	7.2	25_409	0.872
NEB	Nebulin	Muscle - Skeletal	9.9	25_409	0.851
XIRP2	Xin actin-binding repeat-containing protein 2	Muscle - Skeletal	8.0	18_102	0.828
RYR1	Ryanodine receptor 1	Muscle - Skeletal	8.6	29_362	0.827
TMOD4	Tropomodulin-4	Muscle - Skeletal	8.6	30_16	0.827
KLHL40	Kelch-like protein 40	Muscle - Skeletal	7.8	25_409	0.824
SMTNL1	Smoothelin-like protein 1	Muscle - Skeletal	7.2	30_362	0.820
TTN	Titin	Muscle - Skeletal	8.7	18_136	0.810
MYPN	Myopalladin	Muscle - Skeletal	7.3	30_72	0.805
LMOD3	Leiomodin-3	Muscle - Skeletal	7.4	18_136	0.802
MYLPF	Myosin regulatory light chain 2, skeletal muscle isoform	Muscle - Skeletal	10.9	30_161	0.798
LMOD2	Leiomodin-2	Muscle - Skeletal	8.9	30_72	0.798
NRAP	Nebulin-related-anchoring protein	Muscle - Skeletal	10.0	18_136	0.794
MYL2	Myosin regulatory light chain 2, ventricular/cardiac muscle isoform	Muscle - Skeletal	13.7	18_136	0.781
MYLK2	Myosin light chain kinase 2, skeletal/cardiac muscle	Muscle - Skeletal	7.7	30_161	0.777
TNNT1	Troponin T, slow skeletal muscle	Muscle - Skeletal	12.5	18_272	0.776
TNNI1	Troponin I, slow skeletal muscle	Muscle - Skeletal	10.0	30_161	0.773

Gene	Gene name	Tissue with highest expression	Highest median tissue expression (log_2)	Best correlated feature	r
MB	Myoglobin	Muscle - Skeletal	13.1	30_72	0.772
MYH7	Myosin-7	Muscle - Skeletal	12.4	18_136	0.770
KLHL41	Kelch-like protein 41	Muscle - Skeletal	11.6	20_171	0.766
ANP32B	Acidic leucine-rich nuclear phosphoprotein 32 family member B	Nerve - Tibial	8.2	19_468	0.775
CNTF	Ciliary neurotrophic factor	Nerve - Tibial	8.3	29_36	0.769
MXRA8	Matrix remodeling-associated protein 8	Nerve - Tibial	8.5	19_468	0.751
RBPJL	Recombining binding protein suppressor of hairless-like protein	Pancreas	9.5	30_467	0.935
INS	Insulin	Pancreas	10.8	18_158	0.913
CELA2A	Chymotrypsin-like elastase family member 2A	Pancreas	12.6	23_110	0.895
CEL	Bile salt-activated lipase	Pancreas	14.0	23_324	0.866
REG3G	Regenerating islet-derived protein 3-gamma	Pancreas	9.6	30_290	0.815
FSHB	Follitropin subunit beta	Pituitary	7.1	29_279	0.930
POU1F1	Pituitary-specific positive transcription factor 1	Pituitary	7.3	29_279	0.913
GHRHR	Growth hormone-releasing hormone receptor	Pituitary	8.9	29_279	0.900
TSHB	Thyrotropin subunit beta	Pituitary	9.2	29_436	0.869
LHB	Lutropin subunit beta	Pituitary	12.0	30_429	0.844
MYO15A	Unconventional myosin-XV	Pituitary	7.2	29_436	0.803
PRL	Prolactin	Pituitary	15.5	29_436	0.799
TGFBR3L	Transforming growth factor-beta receptor type 3-like protein	Pituitary	8.0	29_436	0.787
GH1	Somatotropin	Pituitary	16.8	30_429	0.763
COL22A1	Collagen alpha-1(XXII) chain	Pituitary	7.2	30_436	0.753
KLK3	Prostate-specific antigen	Prostate	12.8	29_458	0.900
KLK4	Kallikrein-4	Prostate	9.4	29_430	0.821
HOXB13	Homeobox protein Hox-B13	Prostate	7.3	30_458	0.801
KRT77	Keratin, type II cytoskeletal 1b	Skin - Not Sun Exposed	8.6	18_78	0.884
FLG2	Filaggrin-2	Skin - Sun Exposed	9.6	18_78	0.883
LCE2C	Late cornified envelope protein 2C	Skin - Sun Exposed	8.2	18_78	0.878
LCE1A	Late cornified envelope protein 1A	Skin - Sun Exposed	8.9	18_78	0.878
CDSN	Corneodesmosin	Skin - Sun Exposed	8.3	18_78	0.875
LCE2B	Late cornified envelope protein 2B	Skin - Sun Exposed	9.4	18_78	0.874
LCE1B	Late cornified envelope protein 1B	Skin - Sun Exposed	8.0	18_78	0.874
LCE1C	Late cornified envelope protein 1C	Skin - Sun Exposed	9.4	18_78	0.873
LCE6A	Late cornified envelope protein 6A	Skin - Sun Exposed	7.6	18_78	0.870
LCE1F	Late cornified envelope protein 1F	Skin - Sun Exposed	7.6	18_78	0.868
LCE5A	Late cornified envelope protein 5A	Skin - Sun Exposed	7.3	18_78	0.864
LCE2D	Late cornified envelope protein 2D	Skin - Sun Exposed	7.5	18_78	0.863
LCE2A	Late cornified envelope protein 2A	Skin - Sun Exposed	7.2	18_78	0.856
DSC1	Desmocollin-1	Skin - Sun Exposed	8.2	18_78	0.849
KRT2	Keratin, type II cytoskeletal 2 epidermal	Skin - Sun Exposed	12.8	18_78	0.840
FLG	Filaggrin	Skin - Sun Exposed	9.0	18_78	0.835
SERPINB7	Serpin B7	Skin - Sun Exposed	7.4	18_78	0.830
KRT10	Keratin, type I cytoskeletal 10	Skin - Sun Exposed	14.6	18_78	0.826
CASP14	Caspase-14	Skin - Sun Exposed	9.4	18_78	0.820
PSAPL1	Proactivator polypeptide-like 1	Skin - Sun Exposed	7.4	22_265	0.812
CST6	Cystatin-M	Skin - Sun Exposed	9.6	20_329	0.773
ASPRV1	Retroviral-like aspartic protease 1	Skin - Sun Exposed	9.6	18_78	0.772
ALOX12B	Arachidonate 12-lipoxygenase, 12R-type	Skin - Sun Exposed	7.4	25_227	0.757
NR5A1	Steroidogenic factor 1	Spleen; Adrenal Gland	8.0	27_423	0.780
CD19	B-lymphocyte antigen CD19	Spleen	7.9	27_389	0.759
KCNE2	Potassium voltage-gated channel subfamily E member 2	Stomach	8.3	29_287	0.782
DDX4	Probable ATP-dependent RNA helicase DDX4	Testis	7.5	29_39	0.951
DMRTB1	Doublesex- and mab-3-related transcription factor B1	Testis	7.1	29_39	0.949
SHCBP1L	Testicular spindle-associated protein SHCBP1L	Testis	7.7	29_39	0.945
CALR3	Calreticulin-3	Testis	7.1	29_246	0.945
ZPBP2	Zona pellucida-binding protein 2	Testis	7.2	29_39	0.941
SPEM1	Spermatid maturation protein 1	Testis	7.5	29_246	0.939
ACSBG2	Long-chain-fatty-acid--CoA ligase ACSBG2	Testis	8.0	29_39	0.938

Gene	Gene name	Tissue with highest expression	Highest median tissue expression (log_2)	Best correlated feature	r
SPATA19	Spermatogenesis-associated protein 19, mitochondrial	Testis	8.0	29_168	0.937
RNF151	RING finger protein 151	Testis	8.5	29_39	0.937
FSCN3	Fascin-3	Testis	7.2	29_246	0.933
TCP11	T-complex protein 11 homolog	Testis	8.7	29_258	0.930
ODF1	Outer dense fiber protein 1	Testis	9.6	29_39	0.929
IQCF1	IQ domain-containing protein F1	Testis	7.8	29_246	0.926
PRM3	Protamine-3	Testis	7.0	29_246	0.926
TDRG1	Testis development-related protein 1	Testis	7.0	29_39	0.923
SYCP3	Synaptonemal complex protein 3	Testis	7.2	29_39	0.917
CABS1	Calcium-binding and spermatid-specific protein 1	Testis	7.9	29_73	0.915
DKKL1	Dickkopf-like protein 1	Testis	9.3	29_39	0.914
RPL10L	60S ribosomal protein L10-like	Testis	7.5	29_246	0.908
SYCE3	Synaptonemal complex central element protein 3	Testis	7.9	29_39	0.906
CAPZA3	F-actin-capping protein subunit alpha-3	Testis	8.7	29_258	0.906
GTSF1	Gametocyte-specific factor 1	Testis	7.0	29_246	0.905
CCDC42	Coiled-coil domain-containing protein 42	Testis	7.0	29_73	0.903
TXNDC2	Thioredoxin domain-containing protein 2	Testis	7.1	29_246	0.898
TPPP2	Tubulin polymerization-promoting protein family member 2	Testis	8.2	29_73	0.888
AKAP3	A-kinase anchor protein 3	Testis	7.3	29_39	0.880
TNP1	Spermatid nuclear transition protein 1	Testis	13.1	29_39	0.874
TSSK6	Testis-specific serine/threonine-protein kinase 6	Testis	8.2	29_246	0.869
MAEL	Protein maelstrom homolog	Testis	7.3	29_39	0.864
TCP10L	T-complex protein 10A homolog 1	Testis	8.1	29_39	0.864
CCIN	Calicin	Testis	7.3	29_73	0.862
PRAME	Melanoma antigen preferentially expressed in tumors	Testis	7.3	30_39	0.860
PRM1	Sperm protamine P1	Testis	14.3	29_73	0.850
PRM2	Protamine-2	Testis	14.3	29_73	0.849
GGN	Gametogenetin	Testis	7.5	29_39	0.845
ROPN1L	Ropporin-1-like protein	Testis	9.0	29_39	0.844
CABYR	Calcium-binding tyrosine phosphorylation-regulated protein	Testis	8.0	29_39	0.840
INSL3	Insulin-like 3	Testis	8.9	29_39	0.802
ACRBP	Acrosin-binding protein	Testis	9.3	30_246	0.798
PHOSPHO1	Phosphoethanolamine/phosphocholine phosphatase	Testis	7.3	29_246	0.782
SPATA24	Spermatogenesis-associated protein 24	Testis	7.5	29_73	0.778
SPINK2	Serine protease inhibitor Kazal-type 2	Testis	9.2	29_73	0.773
PCSK4	Proprotein convertase subtilisin/kexin type 4	Testis	8.2	30_39	0.773
PRSS21	Testisin	Testis	7.0	30_168	0.773
NKX2-1	Homeobox protein Nkx-2.1	Thyroid	8.5	29_202	0.870
TG	Thyroglobulin	Thyroid	12.3	30_93	0.869
PAX8	Paired box protein Pax-8	Thyroid	10.4	29_499	0.802
FOXE1	Forkhead box protein E1	Thyroid	7.5	27_376	0.778
TRIP6	Thyroid receptor-interacting protein 6	Uterus	7.6	12_209	0.762

Many of the genes significantly correlated with histological features have crucial roles in the development and maintenance of the corresponding tissues. The transcription factors ZIC1, ZIC2, ZIC4, NEUROD1 and NEUROD2 are known to be involved in brain and more specifically cerebellar development [Aruga, 2018; Pieper, 2019], NKX2-5 and BMP10 are implicated in heart development [Akazawa, 2005; Chen, 2004], POU1F1 regulates expression of several genes involved in pituitary development and hormone expression [Kelberman, 2009], NKX2-1, PAX8 and FOXE1 are key thyroid transcription factors, with a fundamental role in the proper formation of the thyroid gland and in maintaining its functional differentiated state in the adult organism [Fernandez, 2015], etc. (Table 6, S3 Table). While a correlational approach such as the present one

cannot replace perturbational tests of causality involved in shaping tissue morphology, it can provide appropriate candidates for such tests.

We also studied the reproducibility of gene-feature correlations w.r.t. the dataset used. More precisely, we determined the Pearson correlation coefficient between the (Fisher transformed) gene-feature correlations computed w.r.t. the validation and respectively test datasets (S3 Fig):

$$r\left(z\left(r\left(\log_2(1+g), f\right)\right)\Big|_{val}, z\left(r\left(\log_2(1+g), f\right)\right)\Big|_{test}\right) = 0.9233$$

$$r\left(r\left(\log_2(1+g), f\right)\Big|_{val}, r\left(\log_2(1+g), f\right)\Big|_{test}\right) = 0.9205$$

where $z(r)$ is the Fisher transform of correlation r. Gene-feature correlations thus show a good reproducibility across datasets.

Visualization of histological features

To investigate the biological relevance of the discovered gene-feature correlations, we analyzed in detail different methods of visualizing the features inferred by convolutional networks. We applied two different visualization methods. The first based on '*guided backpropagation*' determines the regions of a histopathological image that affect the feature of interest most. However, this visualization method cannot be applied exhaustively because it would involve an unmanageably large number of feature-histological image pairs. To deal with this problem, we developed an algorithm for selecting a small number of representative histological image tiles to be visualized with guided backpropagation (S1 File). We also considered a second feature visualization method that is independent of any input image. The method involves generating a synthetic input image that optimizes the network response to the visual feature of interest. Such synthetic images look similar to real histological images that strongly activate the visual feature of interest.

Figures 4-6 show such visualizations of select histological features. Note that *guided backpropagation* (column 2) emphasizes important structural features of the original histological images (column 1). For example, performing guided backpropagation (gBP) of feature 29_499 on a thyroid sample produces patterns that clearly overlap known histological structures of thyroid tissues, namely follicular cells in blue and follicle boundaries in yellow (*row 3*, column 2 of Fig 5). More precisely, the blue "dots" in the gBP image precisely overlap the follicular cells, while the yellow patterns correspond to boundaries of follicles (please compare with the original image from column 1).

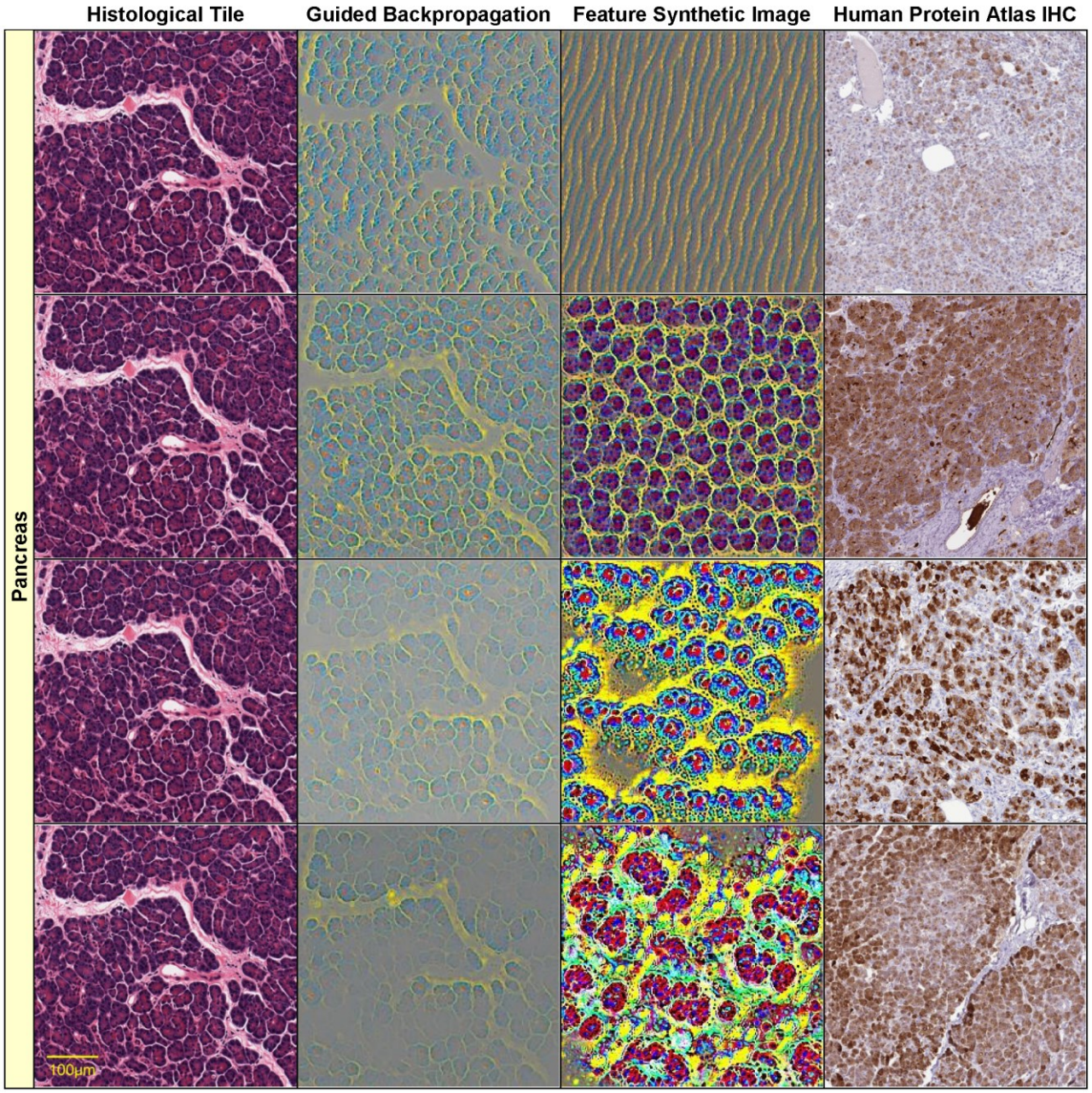


Fig 4. Visualizations of select histological features. The following features (of the form [layer]_[channel]) found correlated with specific genes are visualized on each row:
row 1: 8_55 - CTRC ($r = 0.85$)
row 2: 18_13 - CUZD1 ($r = 0.866$)
row 3: 23_324 - CELA3B ($r = 0.868$)
row 4: 30_467 - AMY2A ($r = 0.928$)
Original image (column 1), guided backpropagation of the feature on the original image (column 2), synthetic image of the feature (column 3), immunohistochemistry image for the corresponding gene from the Human Protein Atlas (column 4). All visualizations are for *pancreas* sample tile GTEX-11NSD-0526_32_5.

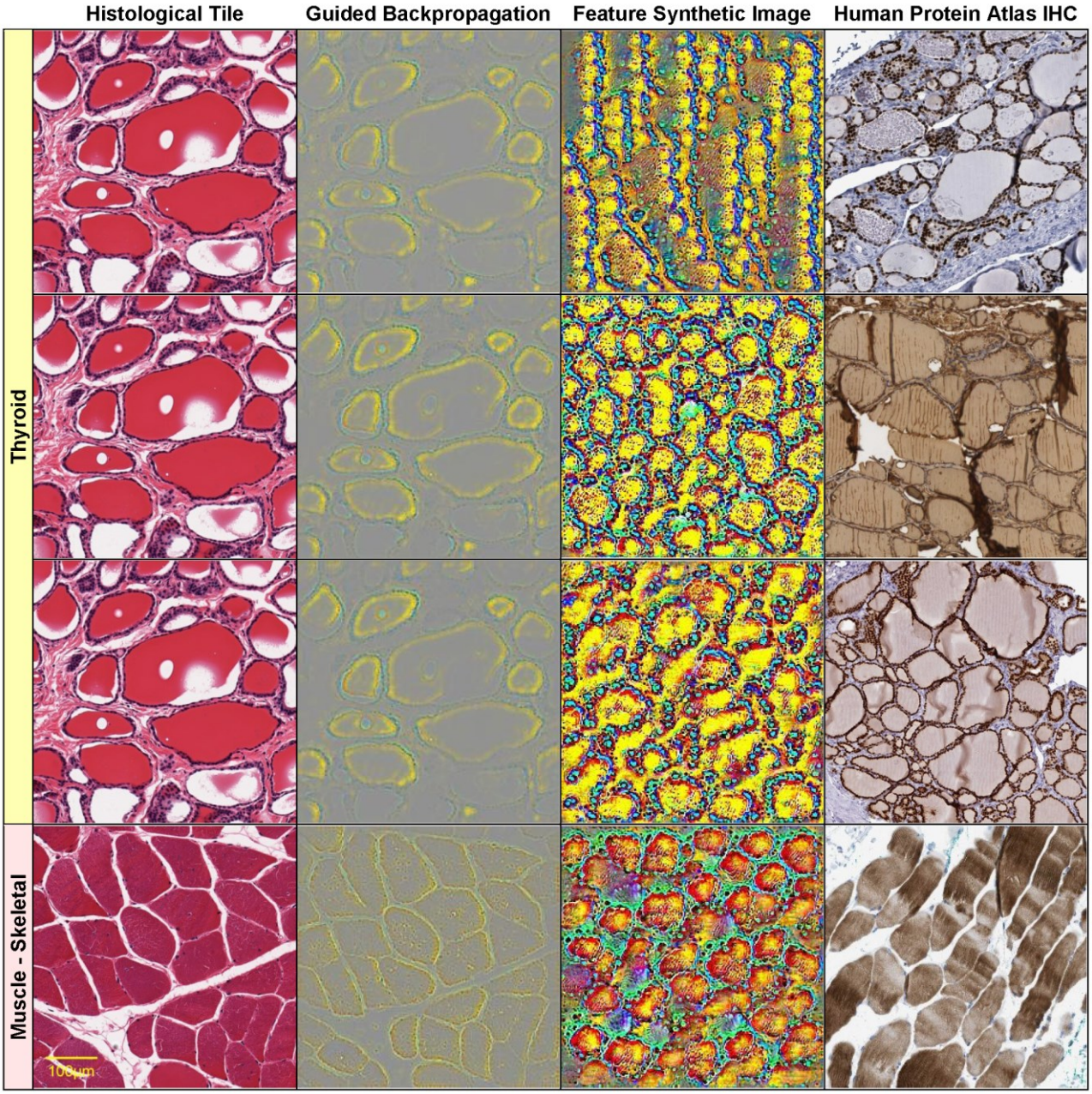


Fig 5. Visualizations of select histological features. The following features (of the form [layer]_[channel]) found correlated with specific genes are visualized on each row:

row 1: 22_405 NKX2-1 ($r = 0.726$) Thyroid GTEX-11NSD-0126_31_16
row 2: 27_343 TG ($r = 0.801$) Thyroid GTEX-11NSD-0126_31_16
row 3: 29_499 PAX8 ($r = 0.802$) Thyroid GTEX-11NSD-0126_31_16
row 4: 25_260 NEB ($r = 0.831$) Muscle - Skeletal GTEX-145ME-2026_39_19

Original image (column 1), guided backpropagation of the feature on the original image (column 2), synthetic image of the feature (column 3), immunohistochemistry image for the corresponding gene from the Human Protein Atlas (column 4).

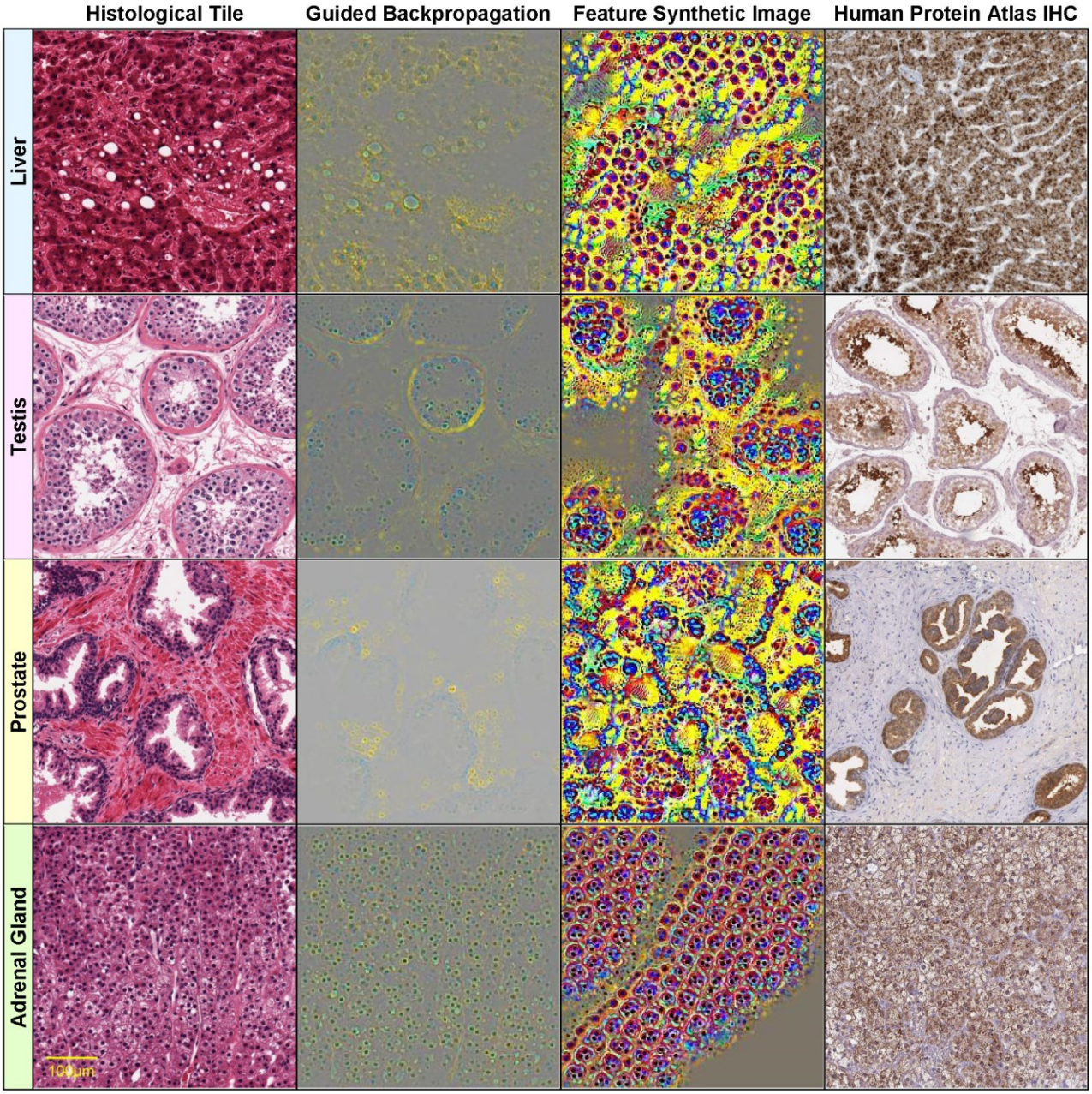


Fig 6. Visualizations of select histological features. The following features (of the form [layer]_[channel]) found correlated with specific genes are visualized on each row:

row 1: 29_234 AGXT (r = 0.939) Liver GTEX-Q2AG-1126_9_7
row 2: 29_118 CALR3 (r = 0.829) Testis GTEX-11NSD-1026_5_27
row 3: 30_244 KLK3 (r = 0.828) Prostate GTEX-V955-1826_8_13
row 4: 20_137 CYP11B1 (r = 0.850) Adrenal Gland GTEX-QLQW-0226_25_9

Original image (column 1), guided backpropagation of the feature on the original image (column 2), synthetic image of the feature (column 3), immunohistochemistry image for the corresponding gene from the Human Protein Atlas (column 4).

On the other hand, column 3 displays a synthetically generated image that maximizes the corresponding feature (29_499). The image was generated by optimization of the feature using incremental changes to an initially random input image. Still, the synthetic image shows a remarkable resemblance to the histological structure of thyroid tissue (except for color differences due to color normalization).

We also show in column 4 the expression of PAX8 in an immunohistochemistry (IHC) stain of thyroid tissue from the Human Protein Atlas [Uhlén, 2015]. Note that PAX8 was identified as a correlate of the above-mentioned feature (29_499) with correlation r=0.802 (Table 6) and is specifically expressed in the follicular cells (corresponding to the blue "dots" in the gBP image from column 3). PAX8 encodes a member of the paired box family of transcription factors and is known to be involved in thyroid follicular cell development and expression of thyroid-specific genes. Although PAX8 is expressed during embryonic development and is subsequently turned off in most adult tissues, its expression is maintained in the thyroid [Fernandez, 2015].

Note that the features with significant gene correlations do not necessarily belong to the final layers of the convolutional neural network (as one may expect the final layers to be better correlated with the training classes, i.e. the tissue types). Fig 4 illustrates this by showing four features *at various layers* (8, 18, 23 and 30) of the VGG16 architecture (which has in total 31 layers, numbered 0 to 30) for a pancreas slide. Lower layers tend to capture simpler visual features, such as blue/yellow edges in the synthetic image of feature 8_55 (row 1, column 3 of Fig 4). In the corresponding guided backpropagation image (column 2), these edges are sufficient to emphasize the yellow boundaries between the blue pancreatic acinar cells. Higher layer features detect more complex histological features, such as round "cell-like" structures that tile the entire image (feature 18_13, row 2, column 3 of Fig 4), or more complex blue cell-like structures (including red nuclei) separated by yellow interstitia (feature 23_324, row 3, column 3 of Fig 4). The highest level feature, 30_467 (row 4, column 3 of Fig 4) seems to detect even more complex patterns involving mostly yellow connective tissue and surrounding red/blue "cells".

Other examples of visualizations of features from different layers are shown in rows 1-3 of Fig 5 for a thyroid slide. The lower level feature 22_405 detects boundaries between blue follicular cells and yellow follicle lumens, while the higher level features 27_343 and 29_499 are activated by more complex, round shapes of blue follicular cells surrounding yellow-red follicle interiors. All of these features are significantly correlated with thyroglobulin TG expression r(TG,22_405)=0.805, r(TG,27_343)=0.801, r(TG,29_499)=0.781, but also with other key thyroidal genes, such as TSHR, NKX2-1, PAX8 and FOXE1: r(TSHR,22_405)=0.828, r(TSHR,27_343)=0.819, r(TSHR,29_499)=0.824, r(FOXE1,22_405)=0.744, r(NKX2-1,22_405)= 0.726, r(NKX2-1,29_499)=0.711, r(PAX8,29_499)=0.802. Note that NKX2-1, FOXE1 and PAX8 are three of the four key thyroid transcription factors, with a fundamental role in the proper formation of the thyroid gland and in maintaining its functional differentiated state in the adult organism [Fernandez, 2015]. The fourth thyroid transcription factor, HHEX, is missing from our analysis due to its expression slightly below the log_2 expression threshold used (the median HHEX expression is 6.84, just below the threshold 7). For illustration purposes, we show, in column 4 of Fig 5, IHC stains for distinct thyroid genes (NKX2-1 for feature 22_405, TG for 27_343 and respectively PAX8 for 29_499). Note that all of these IHC images are remarkably similar to the corresponding guided backpropagation (gBP) images from column 2 and the synthetic images from column 3.

It is remarkable that spatial expression patterns of genes, as assessed by IHC, are frequently very similar to the gBP images of their correlated features. Still, not all significantly correlated gene-

feature pairs display such a good similarity between the gBP image of the feature and the spatial expression pattern of the gene. This is due to the complete lack of spatial specificity of the RNA-seq gene expression data, as well as to the rather coarse-grained classes used for training the visual classifier – just tissue labels, without spatial annotations of tissue substructure. This is the case of tissue-specific genes, which may display *indirect gene-tissue-feature correlations* with distinct spatial specificities of the gene-tissue and respectively feature-tissue correlations. (The gene may be specifically expressed in a certain tissue substructure, while the visual feature may correspond to a distinct substructure of the same tissue. As long as the two different tissue substructures have similar distributions in the tissue slides, we may have indirect gene-tissue-feature correlations without perfect spatial overlap of gene expression and the visual feature.)

For example, the high-level feature 30_467 seems to detect primarily yellow connective tissue in pancreas samples, rather than acinar cells, which express the AMY2A gene (row 4, column 2 of Fig 4). The partial correlation $r(g,f \mid t) = 0.107$ is much lower than $r(g,f) = 0.928$, with a p-value of the conditional independence test p=0.0501 > 0.01. Therefore, the AMY2A expression and the feature 30_467 are independent conditionally on the tissue (i.e. their correlation is indirect via the tissue variable).

To assess the prevalence of such indirect gene(g)-tissue(t)-feature(f) correlations, we performed conditional independence tests using partial correlations of the form $r(g,f \mid t)$ with a significance threshold p=0.01. Table 7 shows the resulting numbers of indirect dependencies for two different significance thresholds of the conditional independence test (α=0.01 and 0.05). The gene-tissue-feature (*g-t-f*) case corresponds to the indirect gene-feature correlations mediated by the tissue variable, discussed above. Indirect feature-gene-tissue (*f-g-t*) dependencies correspond to genes that mediate the feature-tissue correlation (for which the classifier is responsible). There are significantly fewer such indirect dependencies, and even fewer ones of the form *g-f-t*.

Table 7. Numbers of significant gene-feature pairs with indirect dependencies gene-tissue-feature (*g-t-f*), feature-gene-tissue (*f-g-t*), gene-feature-tissue (*g-f-t*). Conditional independence tests with α=0.01 and respectively α=0.05.

Thresholds	*g–f* pairs	*g–t–f*		*f–g–t*		*g–f–t*	
		α=0.01	α=0.05	α=0.01	α=0.05	α=0.01	α=0.05
corr 0.7 log_2 expr 7	27,671	12,442	10,133	5,068	1,486	931	647
corr 0.8 log_2 expr 10	2,176	1,495	1,254	37	12	1	1

The visualizations of the prostate sample from row 3 of Fig 6 illustrate an interesting intermediary case in which although conditioning on the tissue leads to a big drop in correlation (from $r(g,f)$=0.828 to $r(g,f \mid t)$=0.167), the conditional independence test still rejects the null hypothesis (p=0.00215 < 0.01, i.e. there is still conditional *dependence*, although a weak one). The gene-feature correlation is therefore not entirely explainable by the indirect gene-tissue-feature influence. This can also be seen visually in the gBP image of feature 30_244, which detects blue glandular cells and surrounding yellow stroma (row 3, column 2 of Fig 6), similar to the spatial expression pattern of the KLK3 gene, which is specifically expressed in the glandular cells (see IHC image from column 4). The imperfect nature of the dependence is due to the feature detecting mostly basally situated glandular cells (blue) rather than all glandular cells.

Discussion

Due to the limited numbers of high-quality datasets comprising both histopathological images and genomic data for the same subjects, there are very few research publications trying to combine visual features with genomics. Also, the relatively small numbers of samples compared to genes and image features make such integration problems difficult ('small sample problem').

The sparse canonical correlation analysis (CCA) used in [Ash, 2018] is an elegant and very general technique of determining correlations between two *sets* of variables (more precisely, finding linear combinations of variables from each of the two sets that are maximally correlated to one another). Since the sets of genes and respectively image features are very large, CCA is ideal for determining *general trends* (CCA components) in the complicated correlation structure of genes and features. On the other hand, in this paper we are interested in determining *individual* visual features correlated with genes, rather than CCA *composites* of such features, which may be hard to visualize and interpret. Such individual visual features might also be essential for discriminating between visually similar tissues (especially in the case of the number of variables exceeding the number of samples – sparsity constraints mitigate this problem, but do not eliminate it altogether). Moreover, multivariate regression analysis is typically not recommended for small samples. Also, while we use spatially invariant feature encodings obtained by aggregating feature values over both *spatial* and *tile* dimensions, [Ash, 2018] average their feature encodings only over the tile ("window") dimension and not over the spatial dimensions.

Other work searched for relationships between histological images and somatic cancer mutations. For example, [Coudray, 2018] trained a CNN to predict the mutational status of just 10 genes (the most frequently mutated genes in lung cancer, rather than a genome-wide study). Similar mutation predictors were developed for hepatocellular carcinoma [Chen, 2020] and prostate cancer [Schaumberg, 2018]. But the focus of these studies was obtaining mutation predictors, rather than correlating the mutational status with visual histological features.

There is a much larger body of research on purely visual algorithms for analyzing histological slides (without considering correlations with transcriptomics). However, a direct comparison of tissue classification accuracies is probably not very meaningful, since different sets of tissues were used in most papers.

The GTEx dataset was also used to obtain tissue classifiers in [Bizzego, 2019], for 5, 10, 20 and respectively 30 tissues. The highest *tile classification accuracies* reported in [Bizzego, 2019] for 30 tissues are 61.8% for VGG pretrained on Imagenet and respectively 77.1% for the network retrained from scratch on histological images, compared to our 81.5% accuracy for classifying 39 tissues (30 tissue *WSI* classification accuracies are not reported in [Bizzego, 2019]). The dataset in [Bizzego, 2019] contains 53,000 tiles obtained from 787 WSI images from GTEx, for 30 tissues. However, as already mentioned above, some of the additional tissues in our extended set of 39 tissues are hard to distinguish (e.g. 'Artery – Aorta' and 'Artery – Coronary' versus 'Artery – Tibial').

While many researches have already covered normal and diseased tissue classification in digital histopathology, only a small minority have begun to address the problem of interpretability of the resulting models. From this perspective, there are two main contributions of this work, one dealing with the *visual interpretability* of the models, the second involving their *biological interpretability*.

Firstly, we develop *visualization methods* for the features that are inferred automatically by deep learning architectures. These histological features have certain domain-specific characteristics, namely they are *location independent*, as well as precisely *quantifiable*. Quantifiability involves being able to estimate the numbers of occurrences of a specific feature in a given histological image, such as the numbers of specific cellular structures in a slide.

The second main contribution consists in assessing the *biological interpretability* of the deep learning models by correlating their inferred features with matching gene expression data. Such features correlated with gene expression have more than a visual interpretation – they correspond to biological processes at the level of the genome.

It is remarkable that the synthetic images of certain features resemble the corresponding tissue structures extremely well. Genes predominantly expressed in tissues with simpler morphologies tend to be correlated with simpler, lower level visual features, while genes specific to more complex tissue morphologies correlate with higher level features.

However, not all CNN architectures infer features that tend to be well correlated with gene expression profiles, for example VGG networks with batch normalization, Inception_v3 or ResNet (which need batch normalization to deal with their extreme depth). It seems that batch normalization produces slight improvements in tile classification accuracies (though not for *whole* slides) at the expense of biological interpretability (i.e. significantly worse correlations of inferred features with the transcriptome – see Table 4).

Conclusions

Current artificial intelligence systems are still not sufficiently developed to fully take over the tasks of the histopathologist, who is ultimately responsible for the final clinical decision. But the AI system could prove invaluable in assisting the clinical decision process. To do so, it needs to provide the histologist with as much meaningful knowledge as possible. Unfortunately, there is at present no established way to easily explain why a specific decision was made by a network when dealing with a given histopathology image. This is generally unacceptable in the medical community, as clinicians typically need to understand and justify the reasons for a specific decision. A reliable diagnosis must be transparent and fully comprehensible. This is also especially important for obtaining regulatory approval for use in clinical practice [Tizhoosh, 2018]. Of particular concern in the medical field is the uncertainty of the decisions taken by a deep network, which can be radically affected even by the change of very few pixels in an image, in case of a so-called adversarial attack [Athalye, 2018].

Visualizations of the features automatically inferred by a deep neural network are a first step toward making the decisions of the network more *interpretable* and *explainable*. Moreover, correlating the visual histological features with specific gene expression profiles increases the confidence in the *biological interpretability* of these features, since the genes and their expression are responsible for the cell structures that make up these visual features.

This paper deals with identifying transcriptomic correlates of histology using Deep Learning, at first just in normal tissues, as a small step towards bridging the wide explanatory gap between genes, their expression and the complex cellular structures of tissues that make up histological phenotypes. Of course, the relationships discovered in this study are only correlational. For elucidating causality, more complicated, perturbational experiments are needed, but the methodology used here is still applicable.

Acknowledgments

This work was partially supported by the projects PN 1937-0601/2019 and PN 1937-0301/CPN 301 300/2019.

We are also deeply grateful to the GTEx consortium for making the gene expression data and histological images publicly available. The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained from the GTEx Portal (https://gtexportal.org/home/).

Author Contributions

Conceptualization: Liviu Badea

Data curation: Liviu Badea, Emil Stanescu

Formal analysis: Liviu Badea

Investigation: Liviu Badea, Emil Stanescu

Methodology: Liviu Badea

Project administration: Liviu Badea

Software: Liviu Badea

Supervision: Liviu Badea

Validation: Liviu Badea, Emil Stanescu

Visualization: Liviu Badea, Emil Stanescu

Writing – original draft: Liviu Badea

Writing – review & editing: Emil Stanescu

References

1. Xin L, Liu YH, Martin TA, Jiang WG. The era of multigene panels comes? The clinical utility of Oncotype DX and Mammaprint. World journal of oncology. 2017 Apr;8(2):34.
2. Mobadersany P, Yousefi S, Amgad M, Gutman DA, Barnholtz-Sloan JS, Vega JE, Brat DJ, Cooper LA. Predicting cancer outcomes from histology and genomics using convolutional networks. Proceedings of the National Academy of Sciences. 2018 Mar 27;115(13):E2970-9.
3. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. In Advances in neural information processing systems 2014 (pp. 2672-2680).
4. Baak JP, Lindeman J, Overdiep SH, Langley FA. Disagreement of histopathological diagnoses of different pathologists in ovarian tumors—with some theoretical considerations. European Journal of Obstetrics & Gynecology and Reproductive Biology. 1982 Feb 1;13(1):51-5.
5. Yuan Y, Van Allen EM, Omberg L, Wagle N, Amin-Mansour A, Sokolov A, Byers LA, Xu Y, Hess KR, Diao L, Han L. Assessing the clinical utility of cancer genomic and proteomic data across tumor types. Nature biotechnology. 2014 Jul;32(7):644-52.
6. Aeffner F, Zarella MD, Buchbinder N, Bui MM, Goodman MR, Hartman DJ, Lujan GM, Molani MA, Parwani AV, Lillard K, Turner OC. Introduction to digital image analysis in whole-slide imaging: a white paper from the digital pathology association. Journal of pathology informatics. 2019;10.
7. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems 2012 (pp. 1097-1105).
8. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC. Imagenet large scale visual recognition challenge. International journal of computer vision. 2015 Dec 1;115(3):211-52.
9. Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM, Cancer Genome Atlas Research Network. The cancer genome atlas pan-cancer analysis project. Nature genetics. 2013 Sep 26;45(10):1113.
10. Litjens G, Bandi P, Ehteshami Bejnordi B, Geessink O, Balkenhol M, Bult P, Halilovic A, Hermsen M, van de Loo R, Vogels R, Manson QF. 1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. GigaScience. 2018 Jun;7(6):giy065.
11. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, Hasz R, Walters G, Garcia F, Young N, Foster B. The genotype-tissue expression (GTEx) project. Nature genetics. 2013 Jun;45(6):580-5.
12. Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, Moreira AL, Razavian N, Tsirigos A. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nature medicine. 2018 Oct;24(10):1559-67.
13. Wang C, Yang H, Bartz C, Meinel C. Image captioning with deep bidirectional LSTMs. In Proceedings of the 24th ACM international conference on Multimedia 2016 Oct 1 (pp. 988-997).

14. Komura D, Ishikawa S. Machine learning methods for histopathological image analysis. Computational and structural biotechnology journal. 2018 Jan 1;16:34-42.

15. Spanhol FA, Oliveira LS, Petitjean C, Heutte L. Breast cancer histopathological image classification using convolutional neural networks. In 2016 international joint conference on neural networks (IJCNN) 2016 Jul 24 (pp. 2560-2567). IEEE.

16. Sheikhzadeh F, Guillaud M, Ward RK. Automatic labeling of molecular biomarkers of whole slide immunohistochemistry images using fully convolutional networks. arXiv preprint arXiv:1612.09420. 2016 Dec 30.

17. Shah M, Wang D, Rubadue C, Suster D, Beck A. Deep learning assessment of tumor proliferation in breast cancer histological images. In 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 2017 Nov 13 (pp. 600-603). IEEE.

18. Chen H, Qi X, Yu L, Heng PA. DCAN: deep contour-aware networks for accurate gland segmentation. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition 2016 (pp. 2487-2496).

19. Caie PD, Turnbull AK, Farrington SM, Oniscu A, Harrison DJ. Quantification of tumour budding, lymphatic vessel density and invasion through image analysis in colorectal cancer. Journal of translational medicine. 2014 Dec 1;12(1):156.

20. Chen M, Zhang B, Topatana W, Cao J, Zhu H, Juengpanich S, Mao Q, Yu H, Cai X. Classification and mutation prediction based on histopathology H&E images in liver cancer using deep learning. npj Precision Oncology. 2020 Jun 8;4(1):1-7.

21. Schaumberg AJ, Rubin MA, Fuchs TJ. H&E-stained whole slide image deep learning predicts SPOP mutation state in prostate cancer. BioRxiv. 2018 Jan 1:064279.

22. Yu KH, Zhang C, Berry GJ, Altman RB, Ré C, Rubin DL, Snyder M. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. Nature communications. 2016 Aug 16;7(1):1-0.

23. Hochreiter S, Schmidhuber J. Long short-term memory. Neural computation. 1997 Nov 15;9(8):1735-80.

24. Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078. 2014 Jun 3.

25. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. InInternational Conference on Medical image computing and computer-assisted intervention 2015 Oct 5 (pp. 234-241). Springer, Cham.

26. Kamnitsas K, Ledig C, Newcombe VF, Simpson JP, Kane AD, Menon DK, Rueckert D, Glocker B. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Medical image analysis. 2017 Feb 1;36:61-78.

27. Gadermayr, M., Gupta, L., Klinkhammer, B.M., Boor, P. and Merhof, D., 2018. Unsupervisedly Training GANs for Segmenting Digital Pathology with Automatically Generated Annotations. arXiv preprint arXiv:1805.10059.

28. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. 2014 Sep 4.

29. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. InProceedings of the IEEE conference on computer vision and pattern recognition 2016 (pp. 770-778).

30. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition 2016 (pp. 2818-2826).

31. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M, Ghemawat S. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467. 2016 Mar 14.

32. Chollet F. et al. Keras: Deep learning library for theano and tensorflow. https://keras.io/.

33. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A. Pytorch: An imperative style, high-performance deep learning library. In Advances in neural information processing systems 2019 (pp. 8026-8037).

34. Xu Y, Li Y, Shen Z, Wu Z, Gao T, Fan Y, Lai M, Eric I, Chang C. Parallel multiple instance learning for extremely large histopathology image analysis. BMC bioinformatics. 2017 Dec;18(1):1-5.

35. GTEx. The Genotype-Tissue Expression (GTEx) project. (https://gtexportal.org/home/)

36. Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167. 2015 Feb 11.

37. Springenberg JT, Dosovitskiy A, Brox T, Riedmiller M. Striving for simplicity: The all convolutional net. arXiv preprint arXiv:1412.6806. 2014 Dec 21.

38. Olah C, Mordvintsev A, Schubert L. Feature visualization. Distill. 2017 Nov 7;2(11):e7.

39. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA. Gene ontology: tool for the unification of biology. Nature genetics. 2000 May;25(1):25-9.

40. Gene Ontology Consortium. The gene ontology resource: 20 years and still GOing strong. Nucleic acids research. 2019 Jan 8;47(D1):D330-8.

41. Aruga J, Millen KJ. ZIC1 Function in Normal Cerebellar Development and Human Developmental Pathology. Advances in experimental medicine and biology. 2018;1046:249-68.

42. Pieper A, Rudolph S, Wieser GL, Götze T, Mießner H, Yonemasu T, Yan K, Tzvetanova I, Castillo BD, Bode U, Bormuth I. NeuroD2 controls inhibitory circuit formation in the molecular layer of the cerebellum. Scientific reports. 2019 Feb 5;9(1):1-7.

43. Akazawa H, Komuro I. Cardiac transcription factor Csx/Nkx2-5: Its role in cardiac development and diseases. Pharmacology & therapeutics. 2005 Aug 1;107(2):252-68.

44. Chen H, Shi S, Acosta L, Li W, Lu J, Bao S, Chen Z, Yang Z, Schneider MD, Chien KR, Conway SJ. BMP10 is essential for maintaining cardiac growth during murine cardiogenesis. Development. 2004 May 1;131(9):2219-31.

45. Kelberman D, Rizzoti K, Lovell-Badge R, Robinson IC, Dattani MT. Genetic regulation of pituitary gland development in human and mouse. Endocrine reviews. 2009 Dec 1;30(7):790-829.

46. Fernandez LP, Lopez-Marquez A, Santisteban P. Thyroid transcription factors in development, differentiation and disease. Nature Reviews Endocrinology. 2015 Jan;11(1):29-42.

47. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson Å, Kampf C, Sjöstedt E, Asplund A, Olsson I, Edlund K, Lundberg E, Navani S, Szigyarto CA, Odeberg J, Djureinovic D, Takanen JO, Hober S, Alm T, Edqvist PH, Berling H, Tegel H, Mulder J, Rockberg J, Nilsson P, Schwenk JM, Hamsten M, von Feilitzen K, Forsberg M, Persson L, Johansson F, Zwahlen M, von Heijne G, Nielsen J, Pontén F. Tissue-based map of the human proteome. Science. 2015 Jan 23;347 (6220).

48. Ash JT, Darnell G, Munro D, Engelhardt BE. Joint analysis of gene expression levels and histological images identifies genes associated with tissue morphology. bioRxiv. 2018 Jan 1:458711.

49. Bizzego A, Bussola N, Chierici M, Maggio V, Francescatto M, Cima L, Cristoforetti M, Jurman G, Furlanello C. Evaluating reproducibility of AI algorithms in digital pathology with DAPPER. PLoS computational biology. 2019 Mar 27;15(3):e1006269.

50. Tizhoosh HR, Pantanowitz L. Artificial intelligence and digital pathology: challenges and opportunities. Journal of pathology informatics. 2018;9.

51. Athalye A, Engstrom L, Ilyas A, Kwok K. Synthesizing robust adversarial examples. In International Conference on Machine Learning 2018 Jul 3 (pp. 284-293).

Supporting information

S1 Fig. Histograms of features, genes and their correlation. (A) Histogram of feature values. (B) Histogram of log_2 transformed gene expression values $log_2(1+g)$. (C) Histogram of gene-feature correlations (for genes with highest median tissue log_2 expression over 10). (D) Histogram of gene-feature correlations separately for non-negative features (e.g. outputs of ReLU units) and potentially negative features.

S2 Fig. Numbers of correlated genes for individual features and respectively correlated features per gene. (A) Numbers of correlated genes for individual features. Features are sorted in decreasing order of the corresponding numbers of correlated genes. (B) Numbers of correlated features per gene. Genes are sorted in decreasing order of the corresponding numbers of correlated features. Various correlation thresholds are applied (from 0.7 to 0.95).

S3 Fig. Reproducibility of gene-feature correlations between datasets. Scatter plots of gene-feature correlations computed on the validation- and respectively test dataset. (A) scatter plot of *correlations* (R=0.9205). (B) scatter plot of *Fisher transformed* correlations (R=0.9233)

S1 Table. The numbers of slides and respectively tiles for each data set and tissue type.

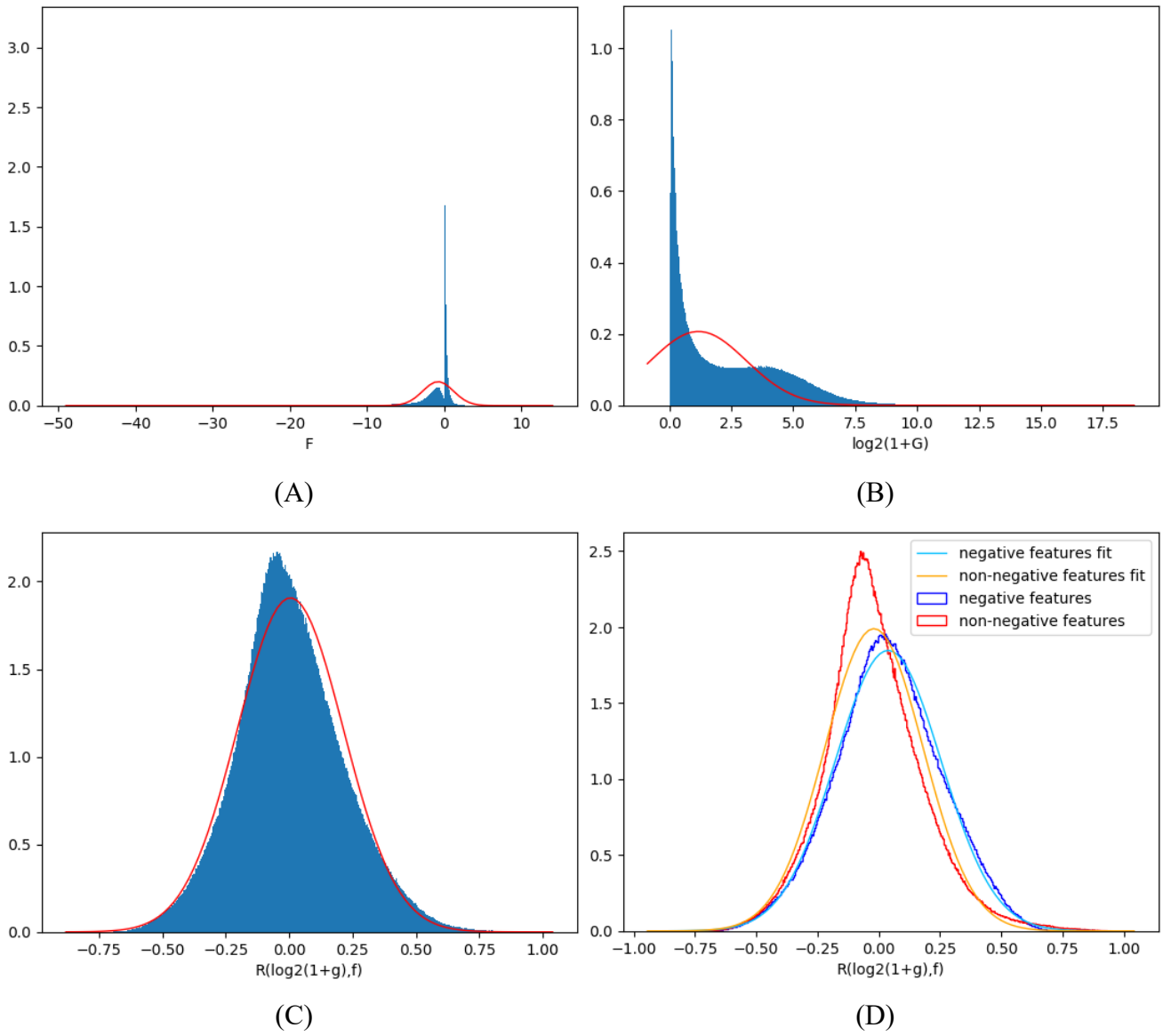
S2 Table. Genes correlated with visual histopathological features. Genes are grouped w.r.t. tissues and ordered by correlation - for each gene we only show the best correlated feature, in the form [layer]_[channel] (for the architecture VGG16 and the test dataset; correlation threshold=0.75, log_2 gene expression threshold=7).

S3 Table. Developmental and transcription regulation genes correlated with visual features. Genes with Gene Ontology '*developmental process*' or '*transcription regulator activity*' annotations are grouped w.r.t. tissues and ordered by correlation - for each gene we only show the best correlated feature, in the form [layer]_[channel] (for the architecture VGG16 and the test dataset; correlation threshold=0.75, log_2 gene expression threshold=7). Genes with '*transcription regulator activity*' are shown in red.

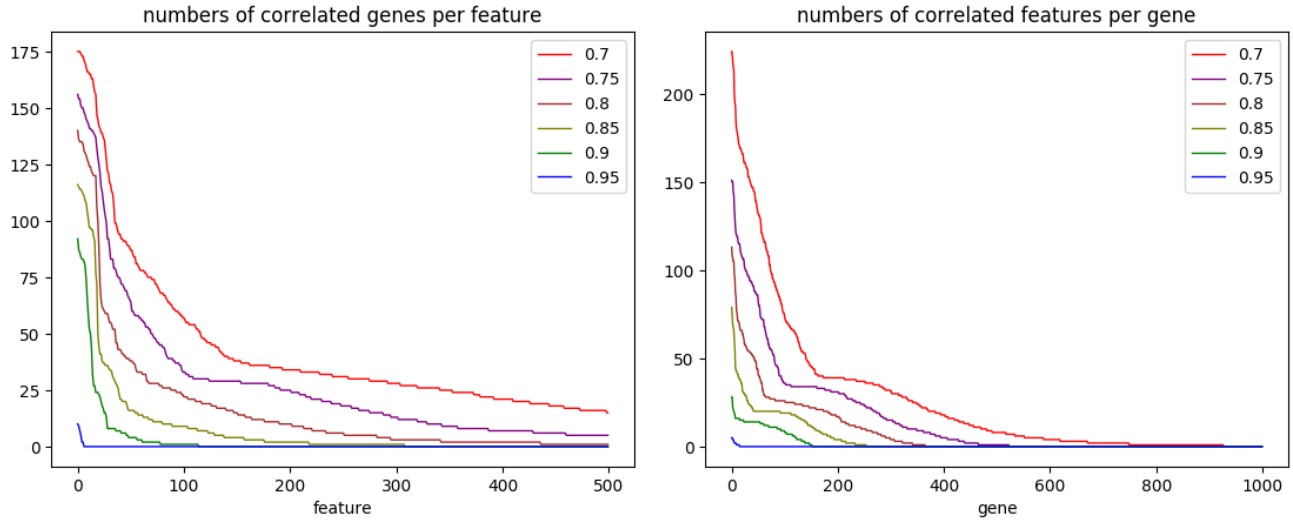
S4 Table. Confusion matrix for the whole slide tissue classifier. VGG16 architecture, test dataset.

S1 File. Supporting information.

S1 Fig. Histograms of features, genes and their correlation. (A) Histogram of feature values. (B) Histogram of log_2 transformed gene expression values $log_2(1+g)$. (C) Histogram of gene-feature correlations (for genes with highest median tissue log_2 expression over 10). (D) Histogram of gene-feature correlations separately for non-negative features (e.g. outputs of ReLU units) and potentially negative features.

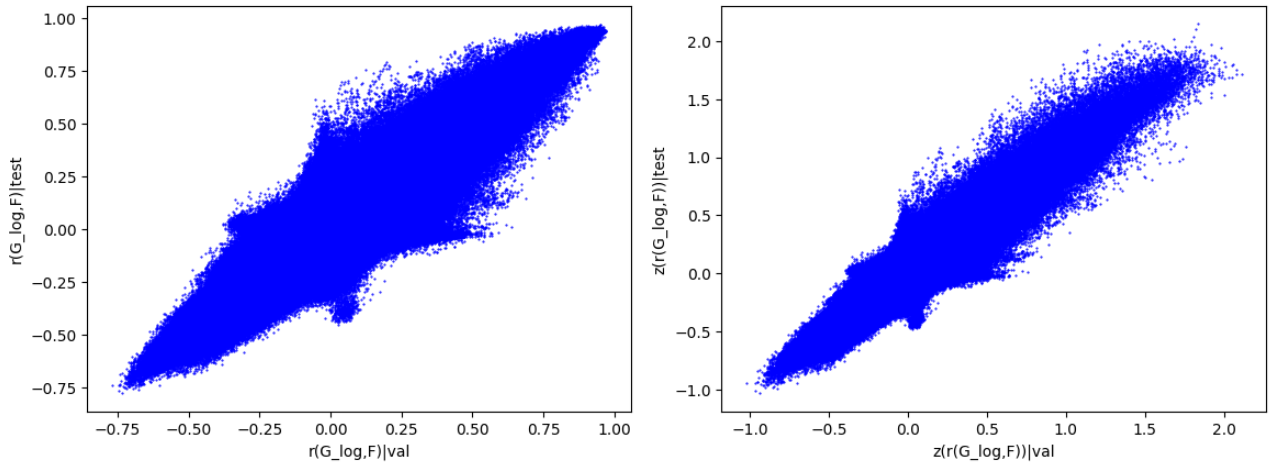


S2 Fig. Numbers of correlated genes for individual features and respectively correlated features per gene. (A) Numbers of correlated genes for individual features. Features are sorted in decreasing order of the corresponding numbers of correlated genes. (B) Numbers of correlated features per gene. Genes are sorted in decreasing order of the corresponding numbers of correlated features. Various correlation thresholds are applied (from 0.7 to 0.95).



(A) (B)

S3 Fig. Reproducibility of gene-feature correlations between datasets. Scatter plots of gene-feature correlations computed on the validation- and respectively test dataset. (A) scatter plot of *correlations* (R=0.9205). (B) scatter plot of *Fisher transformed* correlations (R=0.9233)



(A) (B)

S1 Table. The numbers of slides and respectively tiles for each data set and tissue type.

Tissue	slides				tiles			
	train	*validation*	*test*	*total*	*train*	*validation*	*test*	*total*
Adipose - Subcutaneous	22	7	7	36	6,915	2,203	2,318	11,436
Adipose - Visceral (Omentum)	12	5	4	21	3,374	1,668	1,008	6,050
Adrenal Gland	29	9	9	47	10,007	3,382	2,040	15,429
Artery - Aorta	25	8	9	42	7,346	2,203	2,646	12,195
Artery - Coronary	30	10	10	50	1,559	816	703	3,078
Artery - Tibial	27	9	9	45	1,581	687	529	2,797
Bladder	7	2	2	11	4,472	1,066	1,276	6,814
Brain - Cerebellum	29	10	10	49	7,449	2,848	1,958	12,255
Brain - Cortex	30	9	10	49	8,793	2,495	2,845	14,133
Breast - Mammary Tissue	25	9	8	42	7,417	3,739	3,625	14,781
Cervix - Ectocervix	3	2	1	6	1,161	908	823	2,892
Cervix - Endocervix	3	1	1	5	1,375	221	597	2,193
Colon - Sigmoid	30	10	10	50	12,413	3,078	4,067	19,558
Colon - Transverse	28	9	10	47	9,754	3,698	3,159	16,611
Esophagus - Gastroesophageal Junction	28	10	10	48	13,195	5,032	5,800	24,027
Esophagus - Mucosa	28	10	10	48	6,225	2,086	1,937	10,248
Esophagus - Muscularis	28	8	9	45	12,076	3,246	3,612	18,934
Fallopian Tube	4	1	2	7	638	214	170	1,022
Heart - Atrial Appendage	30	10	10	50	10,691	3,833	4,337	18,861
Heart - Left Ventricle	29	9	9	47	10,861	3,021	2,772	16,654
Kidney - Cortex	27	9	9	45	10,139	3,227	3,365	16,731
Liver	30	10	10	50	13,229	4,111	4,197	21,537
Lung	30	10	9	49	10,314	3,181	3,278	16,773
Minor Salivary Gland	30	10	10	50	4,426	1,527	1,878	7,831
Muscle - Skeletal	30	10	10	50	12,499	3,732	4,059	20,290
Nerve - Tibial	29	9	9	47	3,359	1,365	789	5,513
Ovary	29	10	9	48	12,178	3,475	3,594	19,247
Uterus	29	10	10	49	12,229	4,281	4,025	20,535
Vagina	30	10	10	50	16,274	5,171	5,312	26,757
Pancreas	29	9	10	48	12,863	3,965	4,799	21,627
Pituitary	30	10	9	49	4,940	1,598	1,793	8,331
Prostate	29	9	10	48	13,939	3,718	4,792	22,449
Skin - Not Sun Exposed	30	10	10	50	13,990	4,644	4,128	22,762
Skin - Sun Exposed	30	10	10	50	11,582	4,911	3,545	20,038
Small Intestine - Terminal Ileum	29	10	10	49	10,455	3,158	3,375	16,988
Spleen	30	9	10	49	12,747	4,595	4,530	21,872
Stomach	30	10	9	49	16,753	4,689	3,799	25,241
Testis	30	8	10	48	9,556	2,891	3,627	16,074
Thyroid	28	9	10	47	10,566	3,440	4,918	18,924
Total	1,006	330	334	1,670	349,340	114,123	116,025	579,488

S2 Table. Genes correlated with visual histopathological features. Genes are grouped w.r.t. tissues and ordered by correlation - for each gene we only show the best correlated feature, in the form [layer]_[channel] (for the architecture VGG16 and the test dataset; correlation threshold=0.75, log_2 gene expression threshold=7).

Gene	Gene name	Tissue with highest expression	Highest median tissue expression (log_2)	Best correlated feature	r
FABP4	fatty acid binding protein 4	Adipose - Visceral (Omentum)	12.6	1_14	0.767
CYP11B1	cytochrome P450 family 11 subfamily B member 1	Adrenal Gland	11.7	20_467	0.852
CYP17A1	cytochrome P450 family 17 subfamily A member 1	Adrenal Gland	12.4	20_467	0.805
HSD3B2	hydroxy-delta-5-steroid dehydrogenase, 3 beta- and steroid delta-isomerase 2	Adrenal Gland	11.2	29_123	0.791
CYP21A2	cytochrome P450 family 21 subfamily A member 2	Adrenal Gland	10.7	29_123	0.770
STAR	steroidogenic acute regulatory protein	Adrenal Gland	12.5	29_132	0.770
CYP11A1	cytochrome P450 family 11 subfamily A member 1	Adrenal Gland	9.5	29_148	0.755
TPM4	tropomyosin 4	Artery - Aorta	9.1	19_468	0.767
S100A6	S100 calcium binding protein A6	Artery - Aorta	11.6	19_162	0.760
YAP1	Yes1 associated transcriptional regulator	Artery - Aorta	7.6	19_162	0.758
MARVELD1	MARVEL domain containing 1	Artery - Tibial	7.8	19_162	0.764
PTRF	caveolae associated protein 1	Artery - Tibial	10.1	14_129	0.763
GNG12	G protein subunit gamma 12	Artery - Tibial	7.1	12_209	0.755
GABRA6	gamma-aminobutyric acid type A receptor subunit alpha6	Brain - Cerebellum	7.5	29_447	0.949
ZIC4	Zic family member 4	Brain - Cerebellum	7.0	29_447	0.922
RP11-547D24.1		Brain - Cerebellum	7.2	29_479	0.916
NEUROD2	neuronal differentiation 2	Brain - Cerebellum	7.2	29_463	0.915
GRM4	glutamate metabotropic receptor 4	Brain - Cerebellum	8.3	29_447	0.913
LINC00599	MIR124-1 host gene	Brain - Cerebellum	10.0	29_463	0.907
NEUROD1	neuronal differentiation 1	Brain - Cerebellum	7.8	29_479	0.873
CRTAM	cytotoxic and regulatory T cell molecule	Brain - Cerebellum	7.2	29_479	0.868
KCNJ9	potassium inwardly rectifying channel subfamily J member 9	Brain - Cerebellum	7.1	29_463	0.858
TTC9B	tetratricopeptide repeat domain 9B	Brain - Cerebellum	7.2	27_140	0.856
ZIC2	Zic family member 2	Brain - Cerebellum	7.9	29_463	0.854
GNG13	G protein subunit gamma 13	Brain - Cerebellum	7.2	29_463	0.854
SLC12A5	solute carrier family 12 member 5	Brain - Cerebellum	7.5	27_140	0.850
ZIC1	Zic family member 1	Brain - Cerebellum	8.1	29_479	0.844
TMEM151B	transmembrane protein 151B	Brain - Cerebellum	7.7	27_140	0.844
ELAVL3	ELAV like RNA binding protein 3	Brain - Cerebellum	8.0	29_463	0.834
DIRAS2	DIRAS family GTPase 2	Brain - Cerebellum	7.6	27_140	0.831
C16orf11	proline rich 35	Brain - Cerebellum	7.6	29_447	0.810
GABRD	gamma-aminobutyric acid type A receptor subunit delta	Brain - Cerebellum	9.1	29_463	0.807
GFAP	glial fibrillary acidic protein	Brain - Cerebellum	9.7	29_485	0.804
CACNA1A	calcium voltage-gated channel subunit alpha1 A	Brain - Cerebellum	8.2	27_140	0.794
PVALB	parvalbumin	Brain - Cerebellum	9.1	29_463	0.781
HPCAL4	hippocalcin like 4	Brain - Cerebellum	7.5	29_463	0.781
RP11-981G7.6		Brain - Cerebellum	7.1	30_479	0.770
TMEM145	transmembrane protein 145	Brain - Cerebellum	8.2	25_302	0.766
MAP6D1	MAP6 domain containing 1	Brain - Cerebellum	7.1	29_463	0.765
CPLX2	complexin 2	Brain - Cerebellum	8.7	27_140	0.762
SPOCK2	SPARC (osteonectin), cwcv and kazal like domains proteoglycan 2	Brain - Cerebellum	8.6	27_148	0.761
AGAP2	ArfGAP with GTPase domain, ankyrin repeat and PH domain 2	Brain - Cerebellum	8.0	29_463	0.761
MLC1	modulator of VRAC current 1	Brain - Cerebellum	7.3	29_485	0.760
ATP6V1G2	ATPase H+ transporting V1 subunit G2	Brain - Cerebellum	9.2	29_463	0.752
TUBB4A	tubulin beta 4A class IVa	Brain - Cerebellum	9.5	27_488	0.752
SLC1A2	solute carrier family 1 member 2	Brain - Cortex	8.7	18_8	0.844
SNCB	synuclein beta	Brain - Cortex	9.2	29_485	0.814
GRIN1	glutamate ionotropic receptor NMDA type subunit 1	Brain - Cortex	7.9	18_8	0.812
HPCA	hippocalcin	Brain - Cortex	7.5	29_485	0.811

Gene	Gene name	Tissue with highest expression	Highest median tissue expression (log_2)	Best correlated feature	r
PACSIN1	protein kinase C and casein kinase substrate in neurons 1	Brain - Cortex	8.5	29_463	0.810
SLC17A7	solute carrier family 17 member 7	Brain - Cortex	9.3	18_8	0.799
CAMK2A	calcium/calmodulin dependent protein kinase II alpha	Brain - Cortex	9.2	23_417	0.786
CABP1	calcium binding protein 1	Brain - Cortex	7.3	30_231	0.779
DDN	dendrin	Brain - Cortex	7.9	25_478	0.775
RAPGEF4	Rap guanine nucleotide exchange factor 4	Brain - Cortex	7.6	18_224	0.767
KCNH3	potassium voltage-gated channel subfamily H member 3	Brain - Cortex	7.2	20_380	0.767
CAMKV	CaM kinase like vesicle associated	Brain - Cortex	7.6	18_8	0.766
NSG2	neuronal vesicle trafficking associated 2	Brain - Cortex	7.3	27_278	0.758
CEND1	cell cycle exit and neuronal differentiation 1	Brain - Cortex	8.6	29_485	0.755
KIF5A	kinesin family member 5A	Brain - Cortex	10.0	29_485	0.754
GNG3	G protein subunit gamma 3	Brain - Cortex	8.1	27_278	0.753
SUMO2	small ubiquitin like modifier 2	Cervix - Ectocervix	7.8	19_468	0.771
DDAH2	dimethylarginine dimethylaminohydrolase 2	Cervix - Ectocervix	7.8	19_162	0.765
CYBRD1	cytochrome b reductase 1	Cervix - Ectocervix	8.7	19_162	0.763
NBL1	NBL1, DAN family BMP antagonist	Cervix - Ectocervix	9.6	19_162	0.758
CRTAP	cartilage associated protein	Cervix - Ectocervix	8.1	12_223	0.756
HNRNPUL1	heterogeneous nuclear ribonucleoprotein U like 1	Cervix - Ectocervix	7.2	19_468	0.754
HMGB1	high mobility group box 1	Cervix - Ectocervix	7.4	19_468	0.753
PTMA	prothymosin alpha	Cervix - Endocervix	10.1	19_468	0.802
PIAS3	protein inhibitor of activated STAT 3	Cervix - Endocervix	7.0	19_468	0.764
SPIN1	spindlin 1	Cervix - Endocervix	7.2	19_333	0.752
PLXNB2	plexin B2	Cervix - Endocervix	8.0	12_164	0.750
TACR2	tachykinin receptor 2	Colon - Sigmoid	7.8	23_97	0.753
MANBAL	mannosidase beta like	Fallopian Tube	7.0	21_493	0.752
NKX2-5	NK2 homeobox 5	Heart - Atrial Appendage	7.0	30_214	0.903
BMP10	bone morphogenetic protein 10	Heart - Atrial Appendage	8.3	29_313	0.855
NPPA	natriuretic peptide A	Heart - Atrial Appendage	14.9	29_481	0.777
MYL4	myosin light chain 4	Heart - Atrial Appendage	11.2	29_6	0.755
TECRL	trans-2,3-enoyl-CoA reductase like	Heart - Atrial Appendage	7.1	11_75	0.753
NMRK2	nicotinamide riboside kinase 2	Heart - Atrial Appendage	9.8	30_72	0.751
MYBPC3	myosin binding protein C3	Heart - Left Ventricle	10.8	23_270	0.783
TNNI3	troponin I3, cardiac type	Heart - Left Ventricle	12.0	25_28	0.780
XIRP1	xin actin binding repeat containing 1	Heart - Left Ventricle	7.4	30_72	0.770
TNNT2	troponin T2, cardiac type	Heart - Left Ventricle	11.6	25_31	0.759
CSRP3	cysteine and glycine rich protein 3	Heart - Left Ventricle	9.4	30_72	0.755
AQP2	aquaporin 2	Kidney - Cortex	7.6	29_280	0.890
MIOX	myo-inositol oxygenase	Kidney - Cortex	7.1	29_488	0.888
UMOD	uromodulin	Kidney - Cortex	7.4	27_274	0.874
NAT8	N-acetyltransferase 8 (putative)	Kidney - Cortex	7.2	30_20	0.827
CDH16	cadherin 16	Kidney - Cortex	8.7	30_488	0.816
FXYD4	FXYD domain containing ion transport regulator 4	Kidney - Cortex	7.2	30_280	0.782
F9	coagulation factor IX	Liver	8.1	29_192	0.975
C8A	complement C8 alpha chain	Liver	8.0	29_192	0.973
HAO1	hydroxyacid oxidase 1	Liver	8.0	29_192	0.968
CFHR2	complement factor H related 2	Liver	8.2	29_192	0.966
C9	complement C9	Liver	8.8	29_192	0.962
PLG	plasminogen	Liver	8.7	29_234	0.959
C3P1	complement component 3 precursor pseudogene	Liver	7.8	29_192	0.957
AKR1C4	aldo-keto reductase family 1 member C4	Liver	7.2	29_192	0.956
APOF	apolipoprotein F	Liver	7.2	29_192	0.956
SLC25A47	solute carrier family 25 member 47	Liver	8.6	29_234	0.954
APOA5	apolipoprotein A5	Liver	8.8	29_192	0.954
SERPINC1	serpin family C member 1	Liver	10.1	29_192	0.952
HRG	histidine rich glycoprotein	Liver	9.2	29_234	0.952
F2	coagulation factor II, thrombin	Liver	9.4	29_192	0.949

Gene	Gene name	Tissue with highest expression	Highest median tissue expression (log_2)	Best correlated feature	r
APOC4	apolipoprotein C4	Liver	7.8	29_192	0.949
SERPINA10	serpin family A member 10	Liver	7.6	29_192	0.948
AHSG	alpha 2-HS glycoprotein	Liver	10.7	29_192	0.946
AFM	afamin	Liver	7.9	29_234	0.946
APCS	amyloid P component, serum	Liver	10.7	29_192	0.942
C8B	complement C8 beta chain	Liver	8.2	29_192	0.941
AGXT	alanine--glyoxylate and serine--pyruvate aminotransferase	Liver	10.9	29_234	0.939
CYP8B1	cytochrome P450 family 8 subfamily B member 1	Liver	8.1	29_234	0.938
BAAT	bile acid-CoA:amino acid N-acyltransferase	Liver	7.7	29_192	0.938
ITIH1	inter-alpha-trypsin inhibitor heavy chain 1	Liver	9.1	29_192	0.937
ANGPTL3	angiopoietin like 3	Liver	7.3	29_234	0.934
SLC22A1	solute carrier family 22 member 1	Liver	9.1	30_192	0.933
APOA2	apolipoprotein A2	Liver	12.2	29_192	0.931
CYP4A11	cytochrome P450 family 4 subfamily A member 11	Liver	8.3	29_234	0.924
CPB2	carboxypeptidase B2	Liver	8.5	29_192	0.922
ACSM2A	acyl-CoA synthetase medium chain family member 2A	Liver	7.4	23_292	0.920
PROC	protein C, inactivator of coagulation factors Va and VIIIa	Liver	7.9	29_234	0.919
APOH	apolipoprotein H	Liver	11.7	29_234	0.919
ACSM2B	acyl-CoA synthetase medium chain family member 2B	Liver	7.4	23_292	0.919
SLC22A7	solute carrier family 22 member 7	Liver	7.7	29_234	0.900
CPN2	carboxypeptidase N subunit 2	Liver	7.3	29_234	0.900
KNG1	kininogen 1	Liver	8.6	30_234	0.897
FGB	fibrinogen beta chain	Liver	12.6	29_192	0.896
HPX	hemopexin	Liver	11.1	30_192	0.895
CYP2A6	cytochrome P450 family 2 subfamily A member 6	Liver	9.6	29_192	0.888
ITIH2	inter-alpha-trypsin inhibitor heavy chain 2	Liver	7.7	29_192	0.885
PON1	paraoxonase 1	Liver	7.2	18_279	0.885
TAT	tyrosine aminotransferase	Liver	9.2	29_192	0.884
FGL1	fibrinogen like 1	Liver	10.2	18_231	0.879
SERPINA11	serpin family A member 11	Liver	7.8	29_192	0.877
AMBP	alpha-1-microglobulin/bikunin precursor	Liver	11.7	18_231	0.876
G6PC	glucose-6-phosphatase catalytic subunit	Liver	7.1	29_234	0.872
GC	GC vitamin D binding protein	Liver	9.8	29_192	0.867
PAH	phenylalanine hydroxylase	Liver	8.2	22_418	0.863
SULT2A1	sulfotransferase family 2A member 1	Liver	8.6	29_30	0.863
APOC1P1	apolipoprotein C1 pseudogene 1	Liver	7.0	30_30	0.859
FGA	fibrinogen alpha chain	Liver	12.2	30_192	0.858
ASGR2	asialoglycoprotein receptor 2	Liver	8.8	30_192	0.854
CRP	C-reactive protein	Liver	12.6	29_192	0.852
VTN	vitronectin	Liver	11.2	18_279	0.851
TTR	transthyretin	Liver	11.1	18_231	0.850
SLC2A2	solute carrier family 2 member 2	Liver	7.2	29_234	0.849
SLC13A5	solute carrier family 13 member 5	Liver	8.4	27_275	0.845
BHMT	betaine--homocysteine S-methyltransferase	Liver	8.5	23_292	0.844
APOC3	apolipoprotein C3	Liver	13.0	29_192	0.840
ALB	albumin	Liver	15.1	18_231	0.840
ADH1A	alcohol dehydrogenase 1A (class I), alpha polypeptide	Liver	8.0	29_192	0.835
TDO2	tryptophan 2,3-dioxygenase	Liver	7.7	30_205	0.834
CYP2B6	cytochrome P450 family 2 subfamily B member 6	Liver	7.3	29_234	0.834
SLC27A5	solute carrier family 27 member 5	Liver	8.0	29_192	0.832
C8G	complement C8 gamma chain	Liver	8.7	30_192	0.831
RDH16	retinol dehydrogenase 16	Liver	7.9	30_75	0.830
ORM1	orosomucoid 1	Liver	12.5	30_192	0.829
FGG	fibrinogen gamma chain	Liver	12.2	30_192	0.829
TFR2	transferrin receptor 2	Liver	8.8	27_275	0.828
ORM2	orosomucoid 2	Liver	11.3	30_192	0.826

Gene	Gene name	Tissue with highest expression	Highest median tissue expression (log_2)	Best correlated feature	r
C4BPA	complement component 4 binding protein alpha	Liver	8.4	27_380	0.823
CYP2C8	cytochrome P450 family 2 subfamily C member 8	Liver	9.5	23_492	0.821
CFHR1	complement factor H related 1	Liver	8.1	29_192	0.821
DPYS	dihydropyrimidinase	Liver	7.1	18_249	0.819
UGT2B7	UDP glucuronosyltransferase family 2 member B7	Liver	7.2	23_292	0.819
CYP2E1	cytochrome P450 family 2 subfamily E member 1	Liver	11.1	30_192	0.817
SERPIND1	serpin family D member 1	Liver	7.6	29_192	0.810
SERPINA6	serpin family A member 6	Liver	7.7	29_234	0.806
MAT1A	methionine adenosyltransferase 1A	Liver	9.7	20_471	0.806
F12	coagulation factor XII	Liver	8.7	30_75	0.804
IGFBP1	insulin like growth factor binding protein 1	Liver	7.2	29_75	0.802
ASGR1	asialoglycoprotein receptor 1	Liver	8.7	27_275	0.798
ADH4	alcohol dehydrogenase 4 (class II), pi polypeptide	Liver	9.1	29_192	0.787
APOB	apolipoprotein B	Liver	8.5	30_438	0.784
SLC38A4	solute carrier family 38 member 4	Liver	7.3	29_192	0.780
SAA4	serum amyloid A4, constitutive	Liver	10.4	29_192	0.778
CREB3L3	cAMP responsive element binding protein 3 like 3	Liver	8.2	20_481	0.777
CPS1	carbamoyl-phosphate synthase 1	Liver	8.6	29_192	0.775
SDS	serine dehydratase	Liver	8.9	27_275	0.769
ARG1	arginase 1	Liver	9.0	20_194	0.768
CYP2C9	cytochrome P450 family 2 subfamily C member 9	Liver	8.6	20_481	0.768
A1BG	alpha-1-B glycoprotein	Liver	8.9	30_192	0.768
HPD	4-hydroxyphenylpyruvate dioxygenase	Liver	10.1	30_234	0.761
APOM	apolipoprotein M	Liver	7.2	29_234	0.753
SFTPA1	surfactant protein A1	Lung	10.4	30_155	0.849
SFTPC	surfactant protein C	Lung	9.8	30_155	0.838
SFTPA2	surfactant protein A2	Lung	10.6	30_155	0.803
SCGB3A2	secretoglobin family 3A member 2	Lung	7.7	29_471	0.791
SFTPB	surfactant protein B	Lung	12.2	29_383	0.789
SCGB1A1	secretoglobin family 1A member 1	Lung	9.2	29_155	0.776
NAPSA	napsin A aspartic peptidase	Lung	8.8	30_291	0.774
STATH	statherin	Minor Salivary Gland	7.7	25_200	0.767
IDI2	isopentenyl-diphosphate delta isomerase 2	Muscle - Skeletal	8.0	30_161	0.916
RP11-451G4.2		Muscle - Skeletal	10.3	30_161	0.883
RP11-184I16.4		Muscle - Skeletal	7.2	30_161	0.881
MYF6	myogenic factor 6	Muscle - Skeletal	7.2	25_409	0.872
NEB	nebulin	Muscle - Skeletal	9.9	25_409	0.851
PPP1R27	protein phosphatase 1 regulatory subunit 27	Muscle - Skeletal	9.0	30_161	0.842
CACNG1	calcium voltage-gated channel auxiliary subunit gamma 1	Muscle - Skeletal	7.8	25_409	0.836
XIRP2	xin actin binding repeat containing 2	Muscle - Skeletal	8.0	18_102	0.828
RYR1	ryanodine receptor 1	Muscle - Skeletal	8.6	29_362	0.827
TMOD4	tropomodulin 4	Muscle - Skeletal	8.6	30_16	0.827
KLHL40	kelch like family member 40	Muscle - Skeletal	7.8	25_409	0.824
MYBPC2	myosin binding protein C2	Muscle - Skeletal	9.4	30_161	0.822
SMTNL1	smoothelin like 1	Muscle - Skeletal	7.2	30_362	0.820
TTN	titin	Muscle - Skeletal	8.7	18_136	0.810
MYH1	myosin heavy chain 1	Muscle - Skeletal	10.3	25_409	0.808
MYPN	myopalladin	Muscle - Skeletal	7.3	30_72	0.805
LMOD3	leiomodin 3	Muscle - Skeletal	7.4	18_136	0.802
MYLPF	myosin light chain, phosphorylatable, fast skeletal muscle	Muscle - Skeletal	10.9	30_161	0.798
LMOD2	leiomodin 2	Muscle - Skeletal	8.9	30_72	0.798
NRAP	nebulin related anchoring protein	Muscle - Skeletal	10.0	18_136	0.794
MYL1	myosin light chain 1	Muscle - Skeletal	11.2	30_161	0.789
ATP2A1	ATPase sarcoplasmic/endoplasmic reticulum Ca2+ transporting 1	Muscle - Skeletal	10.0	30_161	0.784
TNNC2	troponin C2, fast skeletal type	Muscle - Skeletal	12.5	30_161	0.782
MYL2	myosin light chain 2	Muscle - Skeletal	13.7	18_136	0.781

Gene	Gene name	Tissue with highest expression	Highest median tissue expression (log_2)	Best correlated feature	r
RP11-357D18.1		Muscle - Skeletal	7.3	27_117	0.779
AMPD1	adenosine monophosphate deaminase 1	Muscle - Skeletal	8.5	30_16	0.779
MYH2	myosin heavy chain 2	Muscle - Skeletal	9.4	30_505	0.779
MYLK2	myosin light chain kinase 2	Muscle - Skeletal	7.7	30_161	0.777
TNNT1	troponin T1, slow skeletal type	Muscle - Skeletal	12.5	18_272	0.776
TRIM63	tripartite motif containing 63	Muscle - Skeletal	9.6	18_136	0.776
TNNI1	troponin I1, slow skeletal type	Muscle - Skeletal	10.0	30_161	0.773
HFE2	hemojuvelin BMP co-receptor	Muscle - Skeletal	7.8	20_305	0.773
MB	myoglobin	Muscle - Skeletal	13.1	30_72	0.772
MYH7	myosin heavy chain 7	Muscle - Skeletal	12.4	18_136	0.770
RPL3L	ribosomal protein L3 like	Muscle - Skeletal	8.2	27_409	0.767
KLHL41	kelch like family member 41	Muscle - Skeletal	11.6	20_171	0.766
COX6A2	cytochrome c oxidase subunit 6A2	Muscle - Skeletal	11.5	30_72	0.759
SLN	sarcolipin	Muscle - Skeletal	10.9	30_72	0.758
ANP32B	acidic nuclear phosphoprotein 32 family member B	Nerve - Tibial	8.2	19_468	0.775
CNTF	ciliary neurotrophic factor	Nerve - Tibial	8.3	29_36	0.769
AHNAK	AHNAK nucleoprotein	Nerve - Tibial	10.0	14_162	0.754
MXRA8	matrix remodeling associated 8	Nerve - Tibial	8.5	19_468	0.751
RPL22	ribosomal protein L22	Ovary	8.9	19_468	0.779
ZFAS1	ZNFX1 antisense RNA 1	Ovary	8.8	19_468	0.775
EEF1A1	eukaryotic translation elongation factor 1 alpha 1	Ovary	12.3	12_164	0.774
RPL3	ribosomal protein L3	Ovary	12.2	12_164	0.765
HNRNPA1	heterogeneous nuclear ribonucleoprotein A1	Ovary	10.0	19_468	0.754
RPL15	ribosomal protein L15	Ovary	9.8	19_468	0.753
CLTA	clathrin light chain A	Ovary	7.5	19_321	0.752
PNLIPRP1	pancreatic lipase related protein 1	Pancreas	11.5	18_425	0.944
SERPINI2	serpin family I member 2	Pancreas	8.0	29_467	0.941
RBPJL	recombination signal binding protein for immunoglobulin kappa J region like	Pancreas	9.5	30_467	0.935
CTRC	chymotrypsin C	Pancreas	12.5	30_467	0.933
RP11-331F4.4		Pancreas	11.6	18_425	0.932
AMY2A	amylase alpha 2A	Pancreas	13.1	30_467	0.928
CELA2B	chymotrypsin like elastase 2B	Pancreas	11.3	30_467	0.917
INS	insulin	Pancreas	10.8	18_158	0.913
SYCN	syncollin	Pancreas	13.8	30_290	0.902
CELA2A	chymotrypsin like elastase 2A	Pancreas	12.6	23_110	0.895
CUZD1	CUB and zona pellucida like domains 1	Pancreas	8.9	22_324	0.895
CTRL	chymotrypsin like	Pancreas	10.8	20_292	0.893
CELP	carboxyl ester lipase pseudogene	Pancreas	8.0	23_324	0.893
CELA3B	chymotrypsin like elastase 3B	Pancreas	13.2	18_425	0.881
CEL	carboxyl ester lipase	Pancreas	14.0	23_324	0.866
CTRB1	chymotrypsinogen B1	Pancreas	14.3	18_158	0.861
PRSS3P2	PRSS3 pseudogene 2	Pancreas	13.4	18_123	0.861
PNLIPRP2	pancreatic lipase related protein 2 (gene/pseudogene)	Pancreas	12.6	18_123	0.851
CPA2	carboxypeptidase A2	Pancreas	13.7	18_123	0.850
PLA2G1B	phospholipase A2 group IB	Pancreas	14.4	18_158	0.846
CTRB2	chymotrypsinogen B2	Pancreas	14.3	18_158	0.845
PRSS3P1	PRSS3 pseudogene 1	Pancreas	13.5	18_123	0.844
PNLIP	pancreatic lipase	Pancreas	16.2	18_123	0.831
GP2	glycoprotein 2	Pancreas	13.6	18_483	0.828
CELA3A	chymotrypsin like elastase 3A	Pancreas	15.0	18_123	0.828
CPB1	carboxypeptidase B1	Pancreas	14.4	18_137	0.822
RP11-133O22.6		Pancreas	8.3	30_290	0.819
CPA1	carboxypeptidase A1	Pancreas	15.9	18_158	0.819

Gene	Gene name	Tissue with highest expression	Highest median tissue expression (log_2)	Best correlated feature	r
REG3G	regenerating family member 3 gamma	Pancreas	9.6	30_290	0.815
CLPS	colipase	Pancreas	16.2	18_123	0.808
RP6-149D17.1		Pancreas	7.8	30_467	0.790
ERP27	endoplasmic reticulum protein 27	Pancreas	8.6	18_377	0.780
PRSS1	serine protease 1	Pancreas	17.0	18_123	0.768
GCG	glucagon	Pancreas	7.8	23_188	0.767
AMY2B	amylase alpha 2B	Pancreas	10.6	22_246	0.754
FSHB	follicle stimulating hormone subunit beta	Pituitary	7.1	29_279	0.930
POU1F1	POU class 1 homeobox 1	Pituitary	7.3	29_279	0.913
GHRHR	growth hormone releasing hormone receptor	Pituitary	8.9	29_279	0.900
TSHB	thyroid stimulating hormone subunit beta	Pituitary	9.2	29_436	0.869
CGA	glycoprotein hormones, alpha polypeptide	Pituitary	13.3	30_429	0.852
LHB	luteinizing hormone subunit beta	Pituitary	12.0	30_429	0.844
OTOS	otospiralin	Pituitary	7.2	29_413	0.831
MYO15A	myosin XVA	Pituitary	7.2	29_436	0.803
PRL	prolactin	Pituitary	15.5	29_436	0.799
POMC	proopiomelanocortin	Pituitary	15.0	29_279	0.794
RP11-566J3.4		Pituitary	7.9	27_217	0.789
TGFBR3L	transforming growth factor beta receptor 3 like	Pituitary	8.0	29_436	0.787
ARHGAP36	Rho GTPase activating protein 36	Pituitary	8.4	29_429	0.786
MIR7-3HG	MIR7-3 host gene	Pituitary	11.3	29_279	0.763
GH1	growth hormone 1	Pituitary	16.8	30_429	0.763
COL22A1	collagen type XXII alpha 1 chain	Pituitary	7.2	30_436	0.753
KLK3	kallikrein related peptidase 3	Prostate	12.8	29_458	0.900
KLK2	kallikrein related peptidase 2	Prostate	11.5	29_193	0.898
KLK4	kallikrein related peptidase 4	Prostate	9.4	29_430	0.821
HOXB13	homeobox B13	Prostate	7.3	30_458	0.801
ACPP	acid phosphatase 3	Prostate	10.5	30_339	0.764
MSMB	microseminoprotein beta	Prostate	10.8	30_458	0.764
KRT77	keratin 77	Skin - Not Sun Exposed	8.6	18_78	0.884
DCD	dermcidin	Skin - Not Sun Exposed	11.4	22_265	0.848
MUCL1	mucin like 1	Skin - Not Sun Exposed	9.7	18_78	0.820
C1orf68	chromosome 1 open reading frame 68	Skin - Sun Exposed	9.6	18_78	0.884
FLG2	filaggrin family member 2	Skin - Sun Exposed	9.6	18_78	0.883
SERPINA12	serpin family A member 12	Skin - Sun Exposed	8.9	18_78	0.880
LCE2C	late cornified envelope 2C	Skin - Sun Exposed	8.2	18_78	0.878
LCE1A	late cornified envelope 1A	Skin - Sun Exposed	8.9	18_78	0.878
CDSN	corneodesmosin	Skin - Sun Exposed	8.3	18_78	0.875
LCE2B	late cornified envelope 2B	Skin - Sun Exposed	9.4	18_78	0.874
LCE1B	late cornified envelope 1B	Skin - Sun Exposed	8.0	18_78	0.874
LCE1C	late cornified envelope 1C	Skin - Sun Exposed	9.4	18_78	0.873
LCE6A	late cornified envelope 6A	Skin - Sun Exposed	7.6	18_78	0.870
LCE1F	late cornified envelope 1F	Skin - Sun Exposed	7.6	18_78	0.868
LCE5A	late cornified envelope 5A	Skin - Sun Exposed	7.3	18_78	0.864
LCE2D	late cornified envelope 2D	Skin - Sun Exposed	7.5	18_78	0.863
LOR	loricrin cornified envelope precursor protein	Skin - Sun Exposed	11.5	18_78	0.857
LCE2A	late cornified envelope 2A	Skin - Sun Exposed	7.2	18_78	0.856
DSC1	desmocollin 1	Skin - Sun Exposed	8.2	18_78	0.849
KRT2	keratin 2	Skin - Sun Exposed	12.8	18_78	0.840
FLG	filaggrin	Skin - Sun Exposed	9.0	18_78	0.835
SERPINB7	serpin family B member 7	Skin - Sun Exposed	7.4	18_78	0.830
KRT10	keratin 10	Skin - Sun Exposed	14.6	18_78	0.826
CASP14	caspase 14	Skin - Sun Exposed	9.4	18_78	0.820
WFDC5	WAP four-disulfide core domain 5	Skin - Sun Exposed	8.7	18_78	0.818
CCL27	C-C motif chemokine ligand 27	Skin - Sun Exposed	7.2	18_78	0.812
PSAPL1	prosaposin like 1	Skin - Sun Exposed	7.4	22_265	0.812

Gene	Gene name	Tissue with highest expression	Highest median tissue expression (log_2)	Best correlated feature	r
THEM5	thioesterase superfamily member 5	Skin - Sun Exposed	7.9	18_78	0.795
KPRP	keratinocyte proline rich protein	Skin - Sun Exposed	9.2	18_78	0.794
PSORS1C2	psoriasis susceptibility 1 candidate 2	Skin - Sun Exposed	7.7	18_78	0.781
CST6	cystatin E/M	Skin - Sun Exposed	9.6	20_329	0.773
ASPRV1	aspartic peptidase retroviral like 1	Skin - Sun Exposed	9.6	18_78	0.772
LY6G6C	lymphocyte antigen 6 family member G6C	Skin - Sun Exposed	9.5	25_227	0.757
ALOX12B	arachidonate 12-lipoxygenase, 12R type	Skin - Sun Exposed	7.4	25_227	0.757
SPINK4	serine peptidase inhibitor Kazal type 4	Small Intestine - Terminal Ileum	7.7	29_390	0.765
CD5L	CD5 molecule like	Spleen	9.2	22_385	0.936
NR5A1	nuclear receptor subfamily 5 group A member 1	Spleen	8.0	27_423	0.780
CD19	CD19 molecule	Spleen	7.9	27_389	0.759
GIF	cobalamin binding intrinsic factor	Stomach	10.8	29_287	0.877
PGA4	pepsinogen A4	Stomach	9.3	29_287	0.872
ATP4B	ATPase H+/K+ transporting subunit beta	Stomach	10.6	29_287	0.865
PGA5	pepsinogen A5	Stomach	8.8	29_287	0.864
PGA3	pepsinogen A3	Stomach	8.5	29_287	0.816
MUC5AC	mucin 5AC, oligomeric mucus/gel-forming	Stomach	10.5	29_287	0.811
TFF2	trefoil factor 2	Stomach	11.2	29_287	0.801
GKN1	gastrokine 1	Stomach	12.2	29_287	0.791
KCNE2	potassium voltage-gated channel subfamily E regulatory subunit 2	Stomach	8.3	29_287	0.782
DDX4	DEAD-box helicase 4	Testis	7.5	29_39	0.951
HDGFL1	HDGF like 1	Testis	7.5	29_39	0.951
DMRTB1	DMRT like family B with proline rich C-terminal 1	Testis	7.1	29_39	0.949
FAM71F1	family with sequence similarity 71 member F1	Testis	7.4	29_39	0.947
SHCBP1L	SHC binding and spindle associated 1 like	Testis	7.7	29_39	0.945
AC114783.1		Testis	7.0	29_73	0.945
CALR3	calreticulin 3	Testis	7.1	29_246	0.945
LYPD4	LY6/PLAUR domain containing 4	Testis	7.8	29_39	0.945
SPACA7	sperm acrosome associated 7	Testis	7.3	29_246	0.945
LINC00917	long intergenic non-protein coding RNA 917	Testis	7.7	29_246	0.944
ANKRD7	ankyrin repeat domain 7	Testis	9.6	29_39	0.944
PRED58		Testis	7.8	29_39	0.943
LYZL2	lysozyme like 2	Testis	7.1	29_73	0.943
IQCF2	IQ motif containing F2	Testis	7.5	29_246	0.943
LINC00202-2	family with sequence similarity 238 member B	Testis	7.6	29_246	0.942
GK2	glycerol kinase 2	Testis	7.1	29_73	0.942
TSPAN16	tetraspanin 16	Testis	7.3	29_246	0.942
CETN1	centrin 1	Testis	8.8	29_39	0.942
TEX33	testis expressed 33	Testis	7.2	29_246	0.942
ZPBP2	zona pellucida binding protein 2	Testis	7.2	29_39	0.941
RP11-71E19.5		Testis	7.1	29_246	0.941
RP11-404K5.2		Testis	7.0	29_39	0.940
PDCL2	phosducin like 2	Testis	7.0	29_168	0.940
DEFB123	defensin beta 123	Testis	7.7	29_39	0.940
TMCO2	transmembrane and coiled-coil domains 2	Testis	7.6	29_246	0.940
TEX37	testis expressed 37	Testis	7.3	29_246	0.940
KIF2B	kinesin family member 2B	Testis	7.3	29_73	0.940
SPERT	chibby family member 2	Testis	7.3	29_73	0.940
CCDC70	coiled-coil domain containing 70	Testis	8.3	29_246	0.940
PPP1R2P9	PPP1R2C family member C	Testis	7.3	29_73	0.939
ACTL9	actin like 9	Testis	8.2	29_246	0.939
SPEM1	spermatid maturation 1	Testis	7.5	29_246	0.939
ACTL7B	actin like 7B	Testis	8.3	29_246	0.938
ZNRF4	zinc and ring finger 4	Testis	8.4	29_246	0.938
ACSBG2	acyl-CoA synthetase bubblegum family member 2	Testis	8.0	29_39	0.938
C17orf74	SPEM family member 2	Testis	8.3	29_246	0.938

Gene	Gene name	Tissue with highest expression	Highest median tissue expression (log_2)	Best correlated feature	r
IRGC	immunity related GTPase cinema	Testis	8.5	29_39	0.937
SPATA19	spermatogenesis associated 19	Testis	8.0	29_168	0.937
TEX38	testis expressed 38	Testis	8.9	29_39	0.937
C10orf62	chromosome 10 open reading frame 62	Testis	9.3	29_39	0.937
ACTRT2	actin related protein T2	Testis	8.8	29_246	0.937
RNF151	ring finger protein 151	Testis	8.5	29_39	0.937
C17orf64	chromosome 17 open reading frame 64	Testis	7.9	29_39	0.937
RP11-314A5.3		Testis	7.6	29_246	0.937
C2orf53	proline rich 30	Testis	8.3	29_73	0.936
PGK2	phosphoglycerate kinase 2	Testis	8.3	29_39	0.936
UBQLN3	ubiquilin 3	Testis	8.4	29_246	0.935
BOD1L2	biorientation of chromosomes in cell division 1 like 2	Testis	9.2	29_39	0.935
FNDC8	fibronectin type III domain containing 8	Testis	7.2	29_246	0.934
SPATA42	spermatogenesis associated 42	Testis	7.8	29_246	0.934
HMGB4	high mobility group box 4	Testis	9.1	29_73	0.934
DEFB119	defensin beta 119	Testis	8.2	29_39	0.933
FSCN3	fascin actin-bundling protein 3	Testis	7.2	29_246	0.933
AKAP4	A-kinase anchoring protein 4	Testis	8.8	29_246	0.933
CCDC54	coiled-coil domain containing 54	Testis	8.2	29_39	0.933
WBSCR28	transmembrane protein 270	Testis	7.8	29_246	0.932
C2orf57	testis expressed 44	Testis	8.5	29_246	0.932
CST9L	cystatin 9 like	Testis	7.0	29_39	0.932
C16orf82	chromosome 16 open reading frame 82	Testis	9.3	29_39	0.932
FAM187B	family with sequence similarity 187 member B	Testis	7.0	29_246	0.932
C17orf50	chromosome 17 open reading frame 50	Testis	7.7	29_39	0.931
C1orf234	testis expressed 46	Testis	7.0	29_73	0.931
ACTL7A	actin like 7A	Testis	9.1	29_73	0.931
SEPT12	septin 12	Testis	7.8	29_39	0.930
SPATA8	spermatogenesis associated 8	Testis	9.6	29_39	0.930
TCP11	t-complex 11	Testis	8.7	29_258	0.930
TUBA3C	tubulin alpha 3c	Testis	10.0	29_39	0.929
GTSF1L	gametocyte specific factor 1 like	Testis	8.9	29_246	0.929
ODF1	outer dense fiber of sperm tails 1	Testis	9.6	29_39	0.929
SMCP	sperm mitochondria associated cysteine rich protein	Testis	10.7	29_39	0.928
GSG1	germ cell associated 1	Testis	8.5	29_39	0.928
IQCF1	IQ motif containing F1	Testis	7.8	29_246	0.926
PRM3	protamine 3	Testis	7.0	29_246	0.926
LINC00238	coiled-coil domain containing 196	Testis	7.2	29_246	0.924
SPATA22	spermatogenesis associated 22	Testis	7.2	29_39	0.923
TDRG1	testis development related 1	Testis	7.0	29_39	0.923
CDIPT-AS1	CDIP transferase opposite strand, pseudogene	Testis	7.9	29_39	0.922
LDHC	lactate dehydrogenase C	Testis	8.4	29_39	0.921
GAPDHS	glyceraldehyde-3-phosphate dehydrogenase, spermatogenic	Testis	8.6	29_73	0.921
FATE1	fetal and adult testis expressed 1	Testis	8.8	29_39	0.920
LELP1	late cornified envelope like proline rich 1	Testis	10.8	29_73	0.917
SYCP3	synaptonemal complex protein 3	Testis	7.2	29_39	0.917
AC010982.1		Testis	8.1	29_258	0.916
C6orf99	chromosome 6 open reading frame 99	Testis	7.3	29_39	0.916
CABS1	calcium binding protein, spermatid associated 1	Testis	7.9	29_73	0.915
EPPIN	epididymal peptidase inhibitor	Testis	7.0	29_39	0.915
DKKL1	dickkopf like acrosomal protein 1	Testis	9.3	29_39	0.914
H1FNT	H1.7 linker histone	Testis	9.0	29_39	0.913
TSACC	TSSK6 activating cochaperone	Testis	9.8	29_168	0.912
C20orf141	chromosome 20 open reading frame 141	Testis	7.9	29_73	0.911
RPL10L	ribosomal protein L10 like	Testis	7.5	29_246	0.908
SYCE3	synaptonemal complex central element protein 3	Testis	7.9	29_39	0.906

Gene	Gene name	Tissue with highest expression	Highest median tissue expression (log_2)	Best correlated feature	r
CAPZA3	capping actin protein of muscle Z-line subunit alpha 3	Testis	8.7	29_258	0.906
GTSF1	gametocyte specific factor 1	Testis	7.0	29_246	0.905
CCDC42	coiled-coil domain containing 42	Testis	7.0	29_73	0.903
TXNDC2	thioredoxin domain containing 2	Testis	7.1	29_246	0.898
UBL4B	ubiquitin like 4B	Testis	8.1	29_39	0.897
TKTL1	transketolase like 1	Testis	7.7	29_39	0.895
SPACA3	sperm acrosome associated 3	Testis	7.7	29_39	0.895
RBAKDN	RBAK downstream neighbor	Testis	10.1	29_258	0.889
TPPP2	tubulin polymerization promoting protein family member 2	Testis	8.2	29_73	0.888
FAM209A	family with sequence similarity 209 member A	Testis	7.0	29_246	0.880
AKAP3	A-kinase anchoring protein 3	Testis	7.3	29_39	0.880
SPTY2D1-AS1	SPTY2D1 opposite strand	Testis	7.0	29_73	0.878
MRVI1-AS1	IRAG1 antisense RNA 1	Testis	7.2	29_246	0.878
TEX40	catsper channel auxiliary subunit zeta	Testis	10.0	29_39	0.877
PFN4	profilin family member 4	Testis	7.5	29_39	0.876
PRND	prion like protein doppel	Testis	7.1	29_39	0.876
TMEM31	transmembrane protein 31	Testis	7.0	29_39	0.876
TNP1	transition protein 1	Testis	13.1	29_39	0.874
TSSK6	testis specific serine kinase 6	Testis	8.2	29_246	0.869
FBXO24	F-box protein 24	Testis	7.1	29_39	0.868
C20orf195	fibronectin type III domain containing 11	Testis	8.0	29_39	0.867
PHF7	PHD finger protein 7	Testis	10.9	29_258	0.867
TEX101	testis expressed 101	Testis	7.7	29_168	0.866
MAEL	maelstrom spermatogenic transposon silencer	Testis	7.3	29_39	0.864
TCP10L	t-complex 10 like	Testis	8.1	29_39	0.864
LEMD1	LEM domain containing 1	Testis	7.3	29_39	0.863
CCIN	calicin	Testis	7.3	29_73	0.862
PRAME	PRAME nuclear receptor transcriptional regulator	Testis	7.3	30_39	0.860
FAM209B	family with sequence similarity 209 member B	Testis	7.4	29_246	0.858
CMTM2	CKLF like MARVEL transmembrane domain containing 2	Testis	10.3	29_246	0.851
SMKR1	small lysine rich protein 1	Testis	8.2	30_39	0.850
PRM1	protamine 1	Testis	14.3	29_73	0.850
PRM2	protamine 2	Testis	14.3	29_73	0.849
GGN	gametogenetin	Testis	7.5	29_39	0.845
DYNLRB2	dynein light chain roadblock-type 2	Testis	7.4	29_39	0.845
ROPN1L	rhophilin associated tail protein 1 like	Testis	9.0	29_39	0.844
TSKS	testis specific serine kinase substrate	Testis	8.3	29_258	0.842
CCDC96	coiled-coil domain containing 96	Testis	7.1	29_39	0.842
CLPB	caseinolytic mitochondrial matrix peptidase chaperone subunit B	Testis	8.2	29_372	0.841
CABYR	calcium binding tyrosine phosphorylation regulated	Testis	8.0	29_39	0.840
RP11-758P17.3		Testis	7.5	29_285	0.832
ZMYND10	zinc finger MYND-type containing 10	Testis	8.5	30_39	0.829
LINC00467	long intergenic non-protein coding RNA 467	Testis	9.5	30_14	0.827
CRISP2	cysteine rich secretory protein 2	Testis	10.0	29_168	0.825
AC004381.6		Testis	8.8	29_168	0.823
ACTL10	actin like 10	Testis	7.2	29_73	0.817
FAM71E1	family with sequence similarity 71 member E1	Testis	8.3	30_39	0.811
CCDC11	cilia and flagella associated protein 53	Testis	7.4	30_39	0.810
GS1-259H13.2		Testis	7.7	30_258	0.809
RP11-496H1.2		Testis	7.4	29_39	0.808
INSL3	insulin like 3	Testis	8.9	29_39	0.802
ADAD2	adenosine deaminase domain containing 2	Testis	7.5	29_285	0.799
ACRBP	acrosin binding protein	Testis	9.3	30_246	0.798
ABHD1	abhydrolase domain containing 1	Testis	7.1	29_39	0.789
COX6B2	cytochrome c oxidase subunit 6B2	Testis	7.3	29_258	0.786

Gene	Gene name	Tissue with highest expression	Highest median tissue expression (log_2)	Best correlated feature	r
SPACA4	sperm acrosome associated 4	Testis	7.9	30_246	0.785
PHOSPHO1	phosphoethanolamine/phosphocholine phosphatase 1	Testis	7.3	29_246	0.782
C1orf194	chromosome 1 open reading frame 194	Testis	8.3	30_39	0.778
SPATA24	spermatogenesis associated 24	Testis	7.5	29_73	0.778
CCT6B	chaperonin containing TCP1 subunit 6B	Testis	7.6	29_258	0.777
SOCS7	suppressor of cytokine signaling 7	Testis	7.6	29_379	0.776
ERICH2	glutamate rich 2	Testis	8.5	29_39	0.773
SPINK2	serine peptidase inhibitor Kazal type 2	Testis	9.2	29_73	0.773
SYCE1	synaptonemal complex central element protein 1	Testis	7.6	30_115	0.773
PCSK4	proprotein convertase subtilisin/kexin type 4	Testis	8.2	30_39	0.773
PRSS21	serine protease 21	Testis	7.0	30_168	0.773
TUBA3E	tubulin alpha 3e	Testis	8.5	30_55	0.765
ACR	acrosin	Testis	7.2	29_246	0.761
MIR202	microRNA 202	Testis	9.6	29_240	0.757
PROCA1	protein interacting with cyclin A1	Testis	8.0	29_39	0.756
ZDHHC19	zinc finger DHHC-type palmitoyltransferase 19	Testis	7.1	29_246	0.755
LRWD1	leucine rich repeats and WD repeat domain containing 1	Testis	9.4	30_246	0.754
TUBA3D	tubulin alpha 3d	Testis	8.7	30_73	0.752
CTD-2182N23.1		Thyroid	9.1	29_93	0.937
TSHR	thyroid stimulating hormone receptor	Thyroid	7.3	30_93	0.907
SLC26A7	solute carrier family 26 member 7	Thyroid	8.8	30_499	0.888
NKX2-1	NK2 homeobox 1	Thyroid	8.5	29_202	0.870
TG	thyroglobulin	Thyroid	12.3	30_93	0.869
SFTA3	surfactant associated 3	Thyroid	7.4	27_346	0.842
IYD	iodotyrosine deiodinase	Thyroid	8.5	29_93	0.836
SLC26A4-AS1	SLC26A4 antisense RNA 1	Thyroid	7.1	30_217	0.806
PAX8	paired box 8	Thyroid	10.4	29_499	0.802
FOXE1	forkhead box E1	Thyroid	7.5	27_376	0.778
PCED1A	PC-esterase domain containing 1A	Thyroid	7.5	26_506	0.751
TRIP6	thyroid hormone receptor interactor 6	Uterus	7.6	12_209	0.762

S3 Table. Developmental and transcription regulation genes correlated with visual features. Genes with Gene Ontology '*developmental process*' or '*transcription regulator activity*' annotations are grouped w.r.t. tissues and ordered by correlation - for each gene we only show the best correlated feature, in the form [layer]_[channel] (for the architecture VGG16 and the test dataset; correlation threshold=0.75, log_2 gene expression threshold=7). Genes with '*transcription regulator activity*' are shown in red.

Gene	Gene name	GO	Tissue with highest expression	Highest median tissue expression (log_2)	Best correlated feature	r
FABP4	Fatty acid-binding protein, adipocyte	white fat cell differentiation; brown fat cell differentiation	Adipose - Visceral (Omentum)	12.6	1_14	0.767
CYP17A1	Steroid 17-alpha-hydroxylase/17,20 lyase	sex differentiation	Adrenal Gland	12.4	20_467	0.805
HSD3B2	3 beta-hydroxysteroid dehydrogenase/Delta 5-->4-isomerase type 2	hippocampus development	Adrenal Gland	11.2	29_123	0.791
STAR	Steroidogenic acute regulatory protein, mitochondrial	brain development	Adrenal Gland	12.5	29_132	0.770
TPM4	Tropomyosin alpha-4 chain	osteoblast differentiation	Artery - Aorta	9.1	19_468	0.767
S100A6	Protein S100-A6	axonogenesis	Artery - Aorta	11.6	19_162	0.760
YAP1	Transcriptional coactivator YAP1	vasculogenesis	Artery - Aorta	7.6	19_162	0.758
MARVELD1	MARVEL domain-containing protein 1	myelination	Artery - Tibial	7.8	19_162	0.764
GNG12	Guanine nucleotide-binding protein G(I)/G(S)/G(O) subunit gamma-12	cerebral cortex development	Artery - Tibial	7.1	12_209	0.755
ZIC4	Zinc finger protein ZIC 4	central nervous system development	Brain - Cerebellum	7.0	29_447	0.922
NEUROD2	Neurogenic differentiation factor 2	cerebellar cortex development; positive regulation of neuron differentiation	Brain - Cerebellum	7.2	29_463	0.915
NEUROD1	Neurogenic differentiation factor 1	cerebellum development; neurogenesis	Brain - Cerebellum	7.8	29_479	0.873
CRTAM	Cytotoxic and regulatory T-cell molecule	regulation of T cell differentiation	Brain - Cerebellum	7.2	29_479	0.868
ZIC2	Zinc finger protein ZIC 2	brain development; cell differentiation	Brain - Cerebellum	7.9	29_463	0.854
SLC12A5	Solute carrier family 12 member 5	dendritic spine development; multicellular organism growth	Brain - Cerebellum	7.5	27_140	0.850
ZIC1	Zinc finger protein ZIC 1	brain development	Brain - Cerebellum	8.1	29_479	0.844
ELAVL3	ELAV-like protein 3	nervous system development; cell differentiation	Brain - Cerebellum	8.0	29_463	0.834
PVALB	Parvalbumin alpha	cochlea development	Brain - Cerebellum	9.1	29_463	0.781
HPCAL4	Hippocalcin-like protein 4	central nervous system development	Brain - Cerebellum	7.5	29_463	0.781
CPLX2	Complexin-2	nervous system development; cell differentiation	Brain - Cerebellum	8.7	27_140	0.762
SPOCK2	Testican-2	synapse assembly; regulation of cell differentiation	Brain - Cerebellum	8.6	27_148	0.761
SLC1A2	Excitatory amino acid transporter 2	telencephalon development; multicellular organism growth	Brain - Cortex	8.7	18_8	0.844
GRIN1	Glutamate receptor ionotropic, NMDA 1	cerebral cortex development	Brain - Cortex	7.9	18_8	0.812
HPCA	Neuron-specific calcium-binding protein hippocalcin	brain development	Brain - Cortex	7.5	29_485	0.811
PACSIN1	Protein kinase C and casein kinase substrate in neurons protein 1	neuron projection morphogenesis; positive regulation of dendrite development	Brain - Cortex	8.5	29_463	0.810
SLC17A7	Vesicular glutamate transporter 1	brain development; synaptic vesicle lumen acidification	Brain - Cortex	9.3	18_8	0.799
CAMK2A	Calcium/calmodulin-dependent protein kinase type II subunit alpha	dendritic spine development; regulation of neuron migration	Brain - Cortex	9.2	23_417	0.786
DDN	Dendrin	DNA-binding transcription factor activity, RNA polymerase II-specific	Brain - Cortex	7.9	25_478	0.775
CEND1	Cell cycle exit and neuronal differentiation protein 1	neuron differentiation	Brain - Cortex	8.6	29_485	0.755
KIF5A	Kinesin heavy chain isoform 5A	axon guidance	Brain - Cortex	10.0	29_485	0.754
NBL1	Neuroblastoma suppressor of tumorigenicity 1	animal organ morphogenesis	Cervix - Ectocervix	9.6	19_162	0.758
CRTAP	Cartilage-associated protein	spermatogenesis	Cervix - Ectocervix	8.1	12_223	0.756
HMGB1	High mobility group protein B1	developmental process	Cervix - Ectocervix	7.4	19_468	0.753
PIAS3	E3 SUMO-protein ligase PIAS3	transcription coregulator activity	Cervix - Endocervix	7.0	19_468	0.764

Gene	Gene name	GO	Tissue with highest expression	Highest median tissue expression (log_2)	Best correlated feature	r
SPIN1	Spindlin-1	multicellular organism development	Cervix - Endocervix	7.2	19_333	0.752
PLXNB2	Plexin-B2	brain development	Cervix - Endocervix	8.0	12_164	0.750
NKX2-5	Homeobox protein Nkx-2.5	atrial cardiac muscle cell development	Heart - Atrial Appendage	7.0	30_214	0.903
BMP10	Bone morphogenetic protein 10	atrial cardiac muscle tissue morphogenesis	Heart - Atrial Appendage	8.3	29_313	0.855
NPPA	Natriuretic peptides A	cell growth involved in cardiac muscle cell development	Heart - Atrial Appendage	14.9	29_481	0.777
NMRK2	Nicotinamide riboside kinase 2	negative regulation of myoblast differentiation	Heart - Atrial Appendage	9.8	30_72	0.751
MYBPC3	Myosin-binding protein C, cardiac-type	heart morphogenesis; ventricular cardiac muscle tissue morphogenesis	Heart - Left Ventricle	10.8	23_270	0.783
TNNI3	Troponin I, cardiac muscle	heart development; ventricular cardiac muscle tissue morphogenesis	Heart - Left Ventricle	12.0	25_28	0.780
TNNT2	Troponin T, cardiac muscle	ventricular cardiac muscle tissue morphogenesis; sarcomere organization	Heart - Left Ventricle	11.6	25_31	0.759
CSRP3	Cysteine and glycine-rich protein 3	cardiac muscle tissue development; cardiac myofibril assembly	Heart - Left Ventricle	9.4	30_72	0.755
AQP2	Aquaporin-2	metanephric collecting duct development	Kidney - Cortex	7.6	29_280	0.890
UMOD	Uromodulin	metanephric ascending thin limb development; metanephric distal convoluted tubule development	Kidney - Cortex	7.4	27_274	0.874
PLG	Plasminogen	tissue regeneration	Liver	8.7	29_234	0.959
APOA5	Apolipoprotein A-V	tissue regeneration	Liver	8.8	29_192	0.954
SERPINC1	Antithrombin-III	lactation	Liver	10.1	29_192	0.952
HRG	Histidine-rich glycoprotein	positive regulation of blood vessel remodeling	Liver	9.2	29_234	0.952
F2	Prothrombin	multicellular organism development; regulation of cell shape	Liver	9.4	29_192	0.949
AHSG	Alpha-2-HS-glycoprotein	skeletal system development	Liver	10.7	29_192	0.946
APCS	Serum amyloid P-component	negative regulation of monocyte differentiation	Liver	10.7	29_192	0.942
BAAT	Bile acid-CoA:amino acid N-acyltransferase	liver development; animal organ regeneration	Liver	7.7	29_192	0.938
ANGPTL3	Angiopoietin-related protein 3	artery morphogenesis	Liver	7.3	29_234	0.934
APOA2	Apolipoprotein A-II	animal organ regeneration	Liver	12.2	29_192	0.931
CYP4A11	Cytochrome P450 4A11	kidney development	Liver	8.3	29_234	0.924
CPB2	Carboxypeptidase B2	liver regeneration; negative regulation of hepatocyte proliferation	Liver	8.5	29_192	0.922
PROC	Vitamin K-dependent protein C	positive regulation of establishment of endothelial barrier	Liver	7.9	29_234	0.919
APOH	Beta-2-glycoprotein 1	negative regulation of angiogenesis	Liver	11.7	29_234	0.919
FGB	Fibrinogen beta chain	positive regulation of substrate adhesion-dependent cell spreading	Liver	12.6	29_192	0.896
FGL1	Fibrinogen-like protein 1	hepatocyte proliferation	Liver	10.2	18_231	0.879
G6PC	Glucose-6-phosphatase	multicellular organism growth	Liver	7.1	29_234	0.872
FGA	Fibrinogen alpha chain	positive regulation of substrate adhesion-dependent cell spreading	Liver	12.2	30_192	0.858
ASGR2	Asialoglycoprotein receptor 2	bone mineralization	Liver	8.8	30_192	0.854
CRP	C-reactive protein	negative regulation of macrophage derived foam cell differentiation	Liver	12.6	29_192	0.852
VTN	Vitronectin	liver regeneration; endodermal cell differentiation	Liver	11.2	18_279	0.851
FGG	Fibrinogen gamma chain	positive regulation of substrate adhesion-dependent cell spreading	Liver	12.2	30_192	0.829
IGFBP1	Insulin-like growth factor-binding protein 1	tissue regeneration	Liver	7.2	29_75	0.802
APOB	Apolipoprotein B-100	post-embryonic development	Liver	8.5	30_438	0.784
CREB3L3	Cyclic AMP-responsive element-binding protein 3-like protein 3	DNA-binding transcription factor activity, RNA polymerase II-specific	Liver	8.2	20_481	0.777
CPS1	Carbamoyl-phosphate synthase [ammonia], mitochondrial	hepatocyte differentiation; midgut development	Liver	8.6	29_192	0.775
ARG1	Arginase-1	liver development	Liver	9.0	20_194	0.768

Gene	Gene name	GO	Tissue with highest expression	Highest median tissue expression (log_2)	Best correlated feature	r
SFTPB	Pulmonary surfactant-associated protein B	animal organ morphogenesis	Lung	12.2	29_383	0.789
SCGB1A1	Uteroglobin	embryo implantation	Lung	9.2	29_155	0.776
STATH	Statherin	hepatocyte differentiation; midgut development	Minor Salivary Gland	7.7	25_200	0.767
MYF6	Myogenic factor 6	skeletal muscle tissue development; muscle cell fate commitment	Muscle - Skeletal	7.2	25_409	0.872
NEB	Nebulin	somatic muscle development; muscle fiber development	Muscle - Skeletal	9.9	25_409	0.851
XIRP2	Xin actin-binding repeat-containing protein 2	cardiac muscle tissue morphogenesis	Muscle - Skeletal	8.0	18_102	0.828
RYR1	Ryanodine receptor 1	skeletal muscle fiber development	Muscle - Skeletal	8.6	29_362	0.827
TMOD4	Tropomodulin-4	myofibril assembly	Muscle - Skeletal	8.6	30_16	0.827
KLHL40	Kelch-like protein 40	skeletal muscle fiber development; skeletal muscle fiber differentiation	Muscle - Skeletal	7.8	25_409	0.824
SMTNL1	Smoothelin-like protein 1	muscle organ morphogenesis	Muscle - Skeletal	7.2	30_362	0.820
TTN	Titin	skeletal muscle thin filament assembly; skeletal muscle myosin thick filament assembly; sarcomere organization	Muscle - Skeletal	8.7	18_136	0.810
MYPN	Myopalladin	sarcomere organization	Muscle - Skeletal	7.3	30_72	0.805
LMOD3	Leiomodin-3	skeletal muscle fiber development; myofibril assembly	Muscle - Skeletal	7.4	18_136	0.802
MYLPF	Myosin regulatory light chain 2, skeletal muscle isoform	skeletal muscle tissue development	Muscle - Skeletal	10.9	30_161	0.798
LMOD2	Leiomodin-2	sarcomere organization; myofibril assembly	Muscle - Skeletal	8.9	30_72	0.798
NRAP	Nebulin-related-anchoring protein	muscle fiber development	Muscle - Skeletal	10.0	18_136	0.794
MYL2	Myosin regulatory light chain 2, ventricular/cardiac muscle isoform	muscle cell fate specification; muscle fiber development	Muscle - Skeletal	13.7	18_136	0.781
MYLK2	Myosin light chain kinase 2, skeletal/cardiac muscle	skeletal muscle satellite cell differentiation; skeletal muscle cell differentiation	Muscle - Skeletal	7.7	30_161	0.777
TNNT1	Troponin T, slow skeletal muscle	sarcomere organization	Muscle - Skeletal	12.5	18_272	0.776
TNNI1	Troponin I, slow skeletal muscle	ventricular cardiac muscle tissue morphogenesis	Muscle - Skeletal	10.0	30_161	0.773
MB	Myoglobin	heart development	Muscle - Skeletal	13.1	30_72	0.772
MYH7	Myosin-7	adult heart development; ventricular cardiac muscle tissue morphogenesis	Muscle - Skeletal	12.4	18_136	0.770
KLHL41	Kelch-like protein 41	skeletal muscle fiber development; myofibril assembly	Muscle - Skeletal	11.6	20_171	0.766
ANP32B	Acidic leucine-rich nuclear phosphoprotein 32 family member B	negative regulation of cell differentiation	Nerve - Tibial	8.2	19_468	0.775
CNTF	Ciliary neurotrophic factor	neuron development; positive regulation of axon regeneration	Nerve - Tibial	8.3	29_36	0.769
MXRA8	Matrix remodeling-associated protein 8	establishment of glial blood-brain barrier	Nerve - Tibial	8.5	19_468	0.751
RBPJL	Recombining binding protein suppressor of hairless-like protein	DNA-binding transcription factor activity, RNA polymerase II-specific	Pancreas	9.5	30_467	0.935
INS	Insulin	positive regulation of cell differentiation	Pancreas	10.8	18_158	0.913
CELA2A	Chymotrypsin-like elastase family member 2A	cornification	Pancreas	12.6	23_110	0.895
CEL	Bile salt-activated lipase	presynaptic membrane assembly	Pancreas	14.0	23_324	0.866
REG3G	Regenerating islet-derived protein 3-gamma	negative regulation of keratinocyte differentiation	Pancreas	9.6	30_290	0.815
FSHB	Follitropin subunit beta	follicle-stimulating hormone signaling pathway	Pituitary	7.1	29_279	0.930
POU1F1	Pituitary-specific positive transcription factor 1	somatotropin secreting cell development	Pituitary	7.3	29_279	0.913
GHRHR	Growth hormone-releasing hormone receptor	somatotropin secreting cell development	Pituitary	8.9	29_279	0.900
TSHB	Thyrotropin subunit beta	anatomical structure morphogenesis	Pituitary	9.2	29_436	0.869
LHB	Lutropin subunit beta	male gonad development	Pituitary	12.0	30_429	0.844
MYO15A	Unconventional myosin-XV	inner ear morphogenesis	Pituitary	7.2	29_436	0.803
PRL	Prolactin	positive regulation of lactation	Pituitary	15.5	29_436	0.799

Gene	Gene name	GO	Tissue with highest expression	Highest median tissue expression (log_2)	Best correlated feature	r
TGFBR3L	Transforming growth factor-beta receptor type 3-like protein	epithelial to mesenchymal transition	Pituitary	8.0	29_436	0.787
GH1	Somatotropin	positive regulation of multicellular organism growth	Pituitary	16.8	30_429	0.763
COL22A1	Collagen alpha-1(XXII) chain	endothelial cell morphogenesis	Pituitary	7.2	30_436	0.753
KLK3	Prostate-specific antigen	negative regulation of angiogenesis	Prostate	12.8	29_458	0.900
KLK4	Kallikrein-4	biomineral tissue development	Prostate	9.4	29_430	0.821
HOXB13	Homeobox protein Hox-B13	prostate epithelial cord arborization involved in prostate glandular acinus morphogenesis; epithelial cell maturation involved in prostate gland development	Prostate	7.3	30_458	0.801
KRT77	Keratin, type II cytoskeletal 1b	keratinization; cornification	Skin - Not Sun Exposed	8.6	18_78	0.884
FLG2	Filaggrin-2	epidermis morphogenesis; establishment of skin barrier	Skin - Sun Exposed	9.6	18_78	0.883
LCE2C	Late cornified envelope protein 2C	keratinization	Skin - Sun Exposed	8.2	18_78	0.878
LCE1A	Late cornified envelope protein 1A	cornification	Skin - Sun Exposed	8.9	18_78	0.878
CDSN	Corneodesmosin	epidermis development; keratinocyte differentiation	Skin - Sun Exposed	8.3	18_78	0.875
LCE2B	Late cornified envelope protein 2B	epidermis development; keratinization	Skin - Sun Exposed	9.4	18_78	0.874
LCE1B	Late cornified envelope protein 1B	keratinization	Skin - Sun Exposed	8.0	18_78	0.874
LCE1C	Late cornified envelope protein 1C	keratinization	Skin - Sun Exposed	9.4	18_78	0.873
LCE6A	Late cornified envelope protein 6A	keratinization	Skin - Sun Exposed	7.6	18_78	0.870
LCE1F	Late cornified envelope protein 1F	keratinization	Skin - Sun Exposed	7.6	18_78	0.868
LCE5A	Late cornified envelope protein 5A	keratinization	Skin - Sun Exposed	7.3	18_78	0.864
LCE2D	Late cornified envelope protein 2D	keratinization	Skin - Sun Exposed	7.5	18_78	0.863
LCE2A	Late cornified envelope protein 2A	keratinization	Skin - Sun Exposed	7.2	18_78	0.856
DSC1	Desmocollin-1	keratinization; cornification	Skin - Sun Exposed	8.2	18_78	0.849
KRT2	Keratin, type II cytoskeletal 2 epidermal	keratinocyte development; epidermis development	Skin - Sun Exposed	12.8	18_78	0.840
FLG	Filaggrin	keratinocyte differentiation; establishment of skin barrier	Skin - Sun Exposed	9.0	18_78	0.835
SERPINB7	Serpin B7	positive regulation of glomerular mesangial cell proliferation	Skin - Sun Exposed	7.4	18_78	0.830
KRT10	Keratin, type I cytoskeletal 10	keratinocyte differentiation; positive regulation of epidermis development	Skin - Sun Exposed	14.6	18_78	0.826
CASP14	Caspase-14	epidermis development; keratinization	Skin - Sun Exposed	9.4	18_78	0.820
PSAPL1	Proactivator polypeptide-like 1	epithelial cell differentiation involved in prostate gland development	Skin - Sun Exposed	7.4	22_265	0.812
CST6	Cystatin-M	epidermis development; anatomical structure morphogenesis	Skin - Sun Exposed	9.6	20_329	0.773
ASPRV1	Retroviral-like aspartic protease 1	skin development	Skin - Sun Exposed	9.6	18_78	0.772
ALOX12B	Arachidonate 12-lipoxygenase, 12R-type	establishment of skin barrier	Skin - Sun Exposed	7.4	25_227	0.757
NR5A1	Steroidogenic factor 1	DNA-binding transcription factor activity, RNA polymerase II-specific; adrenal gland development	Spleen; Adrenal Gland	8.0	27_423	0.780
CD19	B-lymphocyte antigen CD19	B-1 B cell differentiation	Spleen	7.9	27_389	0.759
KCNE2	Potassium voltage-gated channel subfamily E member 2	tongue development	Stomach	8.3	29_287	0.782
DDX4	Probable ATP-dependent RNA helicase DDX4	spermatogenesis; cell differentiation	Testis	7.5	29_39	0.951
DMRTB1	Doublesex- and mab-3-related transcription factor B1	DNA-binding transcription factor activity, RNA polymerase II-specific	Testis	7.1	29_39	0.949
SHCBP1L	Testicular spindle-associated protein SHCBP1L	spermatogenesis; cell differentiation	Testis	7.7	29_39	0.945
CALR3	Calreticulin-3	spermatogenesis; cell differentiation	Testis	7.1	29_246	0.945
ZPBP2	Zona pellucida-binding protein 2	acrosome assembly	Testis	7.2	29_39	0.941
SPEM1	Spermatid maturation protein 1	sperm individualization; spermatogenesis	Testis	7.5	29_246	0.939
ACSBG2	Long-chain-fatty-acid--CoA ligase ACSBG2	spermatogenesis; cell differentiation	Testis	8.0	29_39	0.938
SPATA19	Spermatogenesis-associated protein 19, mitochondrial	spermatogenesis; cell differentiation	Testis	8.0	29_168	0.937

Gene	Gene name	GO	Tissue with highest expression	Highest median tissue expression (log_2)	Best correlated feature	r
RNF151	RING finger protein 151	spermatogenesis; cell differentiation	Testis	8.5	29_39	0.937
FSCN3	Fascin-3	spermatid development	Testis	7.2	29_246	0.933
TCP11	T-complex protein 11 homolog	spermatogenesis; cell differentiation	Testis	8.7	29_258	0.930
ODF1	Outer dense fiber protein 1	spermatogenesis; cell differentiation	Testis	9.6	29_39	0.929
IQCF1	IQ domain-containing protein F1	positive regulation of flagellated sperm motility involved in capacitation	Testis	7.8	29_246	0.926
PRM3	Protamine-3	spermatogenesis; cell differentiation	Testis	7.0	29_246	0.926
TDRG1	Testis development-related protein 1	multicellular organism development	Testis	7.0	29_39	0.923
SYCP3	Synaptonemal complex protein 3	spermatogenesis; spermatid development	Testis	7.2	29_39	0.917
CABS1	Calcium-binding and spermatid-specific protein 1	spermatogenesis	Testis	7.9	29_73	0.915
DKKL1	Dickkopf-like protein 1	anatomical structure morphogenesis	Testis	9.3	29_39	0.914
RPL10L	60S ribosomal protein L10-like	spermatogenesis	Testis	7.5	29_246	0.908
SYCE3	Synaptonemal complex central element protein 3	spermatogenesis	Testis	7.9	29_39	0.906
CAPZA3	F-actin-capping protein subunit alpha-3	spermatid development	Testis	8.7	29_258	0.906
GTSF1	Gametocyte-specific factor 1	spermatogenesis; cell differentiation	Testis	7.0	29_246	0.905
CCDC42	Coiled-coil domain-containing protein 42	spermatid development	Testis	7.0	29_73	0.903
TXNDC2	Thioredoxin domain-containing protein 2	spermatogenesis; cell differentiation	Testis	7.1	29_246	0.898
TPPP2	Tubulin polymerization-promoting protein family member 2	spermatogenesis; cell differentiation	Testis	8.2	29_73	0.888
AKAP3	A-kinase anchor protein 3	blastocyst hatching	Testis	7.3	29_39	0.880
TNP1	Spermatid nuclear transition protein 1	spermatid development; spermatogenesis	Testis	13.1	29_39	0.874
TSSK6	Testis-specific serine/threonine-protein kinase 6	multicellular organism development; sperm chromatin condensation	Testis	8.2	29_246	0.869
MAEL	Protein maelstrom homolog	spermatogenesis	Testis	7.3	29_39	0.864
TCP10L	T-complex protein 10A homolog 1	transcription corepressor activity	Testis	8.1	29_39	0.864
CCIN	Calicin	spermatogenesis	Testis	7.3	29_73	0.862
PRAME	Melanoma antigen preferentially expressed in tumors	cell differentiation	Testis	7.3	30_39	0.860
PRM1	Sperm protamine P1	spermatogenesis; cell differentiation	Testis	14.3	29_73	0.850
PRM2	Protamine-2	spermatogenesis; spermatid development	Testis	14.3	29_73	0.849
GGN	Gametogenetin	spermatogenesis; cell differentiation	Testis	7.5	29_39	0.845
ROPN1L	Ropporin-1-like protein	sperm capacitation	Testis	9.0	29_39	0.844
CABYR	Calcium-binding tyrosine phosphorylation-regulated protein	sperm capacitation	Testis	8.0	29_39	0.840
INSL3	Insulin-like 3	spermatogenesis	Testis	8.9	29_39	0.802
ACRBP	Acrosin-binding protein	spermatid development; acrosome assembly	Testis	9.3	30_246	0.798
PHOSPHO1	Phosphoethanolamine/phosphocholine phosphatase	regulation of bone mineralization	Testis	7.3	29_246	0.782
SPATA24	Spermatogenesis-associated protein 24	spermatogenesis; cell differentiation	Testis	7.5	29_73	0.778
SPINK2	Serine protease inhibitor Kazal-type 2	spermatid development; acrosome assembly	Testis	9.2	29_73	0.773
PCSK4	Proprotein convertase subtilisin/kexin type 4	sperm capacitation	Testis	8.2	30_39	0.773
PRSS21	Testisin	spermatogenesis	Testis	7.0	30_168	0.773
NKX2-1	Homeobox protein Nkx-2.1	thyroid gland development	Thyroid	8.5	29_202	0.870
TG	Thyroglobulin	thyroid gland development	Thyroid	12.3	30_93	0.869
PAX8	Paired box protein Pax-8	thyroid gland development	Thyroid	10.4	29_499	0.802
FOXE1	Forkhead box protein E1	thyroid gland development	Thyroid	7.5	27_376	0.778
TRIP6	Thyroid receptor-interacting protein 6	chordate embryonic development	Uterus	7.6	12_209	0.762

S4 Table. Confusion matrix for the whole slide tissue classifier. VGG16 architecture, test dataset. The rows correspond to the ground truth, the columns to the predicted classes. Similar tissues have been marked with colors, e.g. (Artery - Aorta, Artery - Coronary, Artery - Tibial).

	Adipose - Subcutaneous	Adipose - Visceral (Omentum)	Adrenal Gland	Artery - Aorta	Artery - Coronary	Artery - Tibial	Bladder	Brain - Cerebellum	Brain - Cortex	Breast - Mammary Tissue	Cervix - Ectocervix	Cervix - Endocervix	Colon - Sigmoid	Colon - Transverse	Esophagus - Gastroesophageal Junction	Esophagus - Mucosa	Esophagus - Muscularis	Fallopian Tube	Heart - Atrial Appendage	Heart - Left Ventricle	Kidney - Cortex	Liver	Lung	Minor Salivary Gland	Muscle - Skeletal	Nerve - Tibial	Ovary	Pancreas	Pituitary	Prostate	Skin - Not Sun Exposed	Skin - Sun Exposed	Small Intestine - Terminal Ileum	Spleen	Stomach	Testis	Thyroid	Uterus	Vagina	**Total**
Adipose - Subcutaneous	6	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**7**
Adipose - Visceral (Omentum)	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**4**
Adrenal Gland	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**9**
Artery - Aorta	0	0	0	8	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**9**
Artery - Coronary	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**10**
Artery - Tibial	0	0	0	0	1	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**9**
Bladder	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**2**
Brain - Cerebellum	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**10**
Brain - Cortex	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**10**
Breast - Mammary Tissue	1	0	0	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**8**
Cervix - Ectocervix	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	**1**
Cervix - Endocervix	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**
Colon - Sigmoid	0	0	0	0	0	0	0	0	0	0	0	0	9	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**10**
Colon - Transverse	0	0	0	0	0	0	0	0	0	0	0	0	2	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	**10**
Esophagus - Gastroesophageal Junction	0	0	0	0	0	0	0	0	0	0	0	0	1	0	7	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**10**
Esophagus - Mucosa	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**10**
Esophagus - Muscularis	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**9**
Fallopian Tube	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**2**
Heart - Atrial Appendage	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**10**
Heart - Left Ventricle	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**9**
Kidney - Cortex	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**9**
Liver	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**10**
Lung	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**9**
Minor Salivary Gland	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**10**
Muscle - Skeletal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**10**
Nerve - Tibial	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	**9**
Ovary	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	**9**
Pancreas	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	**10**
Pituitary	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0	**9**
Prostate	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	**10**
Skin - Not Sun Exposed (Suprapubic)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	2	0	0	0	0	0	0	0	**10**
Skin - Sun Exposed (Lower leg)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	9	0	0	0	0	0	0	0	**10**
Small Intestine - Terminal Ileum	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	**10**
Spleen	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	**10**
Stomach	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	**9**
Testis	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	**10**
Thyroid	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	**10**
Uterus	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	**10**
Vagina	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	**10**
Total	**8**	**4**	**9**	**8**	**11**	**9**	**2**	**10**	**10**	**8**	**0**	**1**	**12**	**8**	**10**	**10**	**8**	**2**	**9**	**10**	**9**	**11**	**8**	**10**	**10**	**8**	**9**	**10**	**9**	**10**	**9**	**11**	**10**	**10**	**10**	**10**	**10**	**10**	**11**	**334**

S1 File. Supporting information.

Selection of representative whole slide image tiles

Unfortunately, guided backpropagation visualization cannot be applied exhaustively because it would involve an unmanageably large number of feature-histological image (f,X) pairs. To deal with this problem, we developed an algorithm for selecting a small number of representative histological image tiles to be visualized with guided backpropagation.

select k=3 representative samples for each tissue:

consider the (gene, feature, tissue) table (g,f,t) with corr(g,f) ≥ 0.8 and highest median tissue log_2 gene expression ≥ 10

for each tissue t

get the samples of the given tissue: samples(t)

get the features f such that (f,t) appear in the table

construct the matrix val(f,s) = the value of the feature f in sample s, for s in samples(t) and f such that (f,t) in table

normalize the rows of this matrix: val(f,s) = val(f,s) / $\text{norm}_{s'}$ val(f,s')

select the samples with the k largest values of $\min_f$ val(f,s) as representative samples for tissue t: selected_samples(t) = k-argmax_s $\min_f$ val(f,s)

Note that $\min_f$ val(f,s) ensures that if the sample s is chosen as representative for tissue t, then all features f associated to this tissue have a value at least $\min_f$ val(f,s).

The same algorithm is applied for selecting a single representative tile from each sample.

Guided backpropagation image normalization

Guided backpropagation generates images of gradients X that must be normalized for visualization. We use the following log_2 transformation of the gradients for better emphasizing the lower intensity details:

$$X_{\log} = \log_2\left(1+\lambda\frac{X^+}{M}\right) - \log_2\left(1+\lambda\frac{X^-}{M}\right)$$

where $X^{\pm} = \frac{|X| \pm X}{2}$, $M = \max\left(\max(X^+), \max(X^-)\right)$, λ=10. In turn, $X_{\log}$ is further normalized to the range [0,1]:

$$X_{norm} = \frac{X_{\log} - \min(X_{\log})}{\max\left(X_{\log} - \min(X_{\log})\right)}.$$