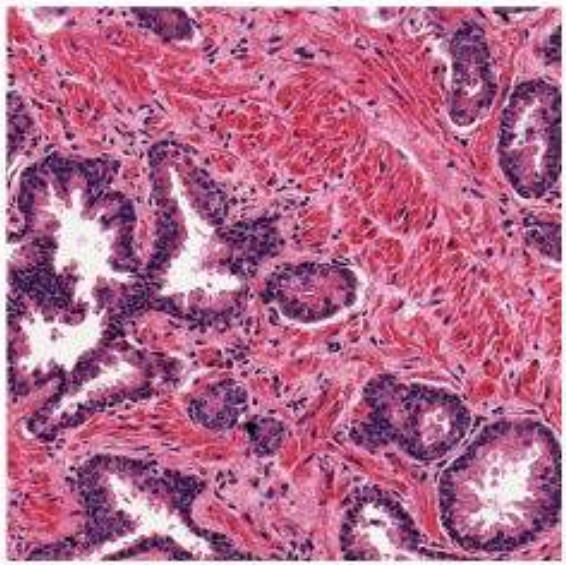
Intersecting pathology images and gene expression data to understand drivers of complex phenotypes

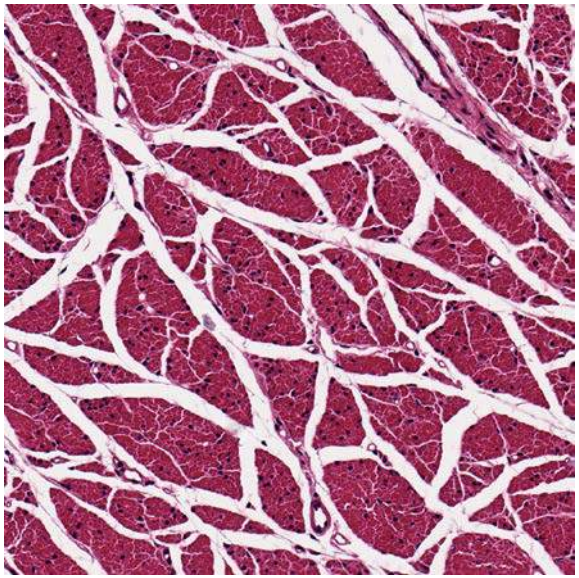


Barbara E Engelhardt

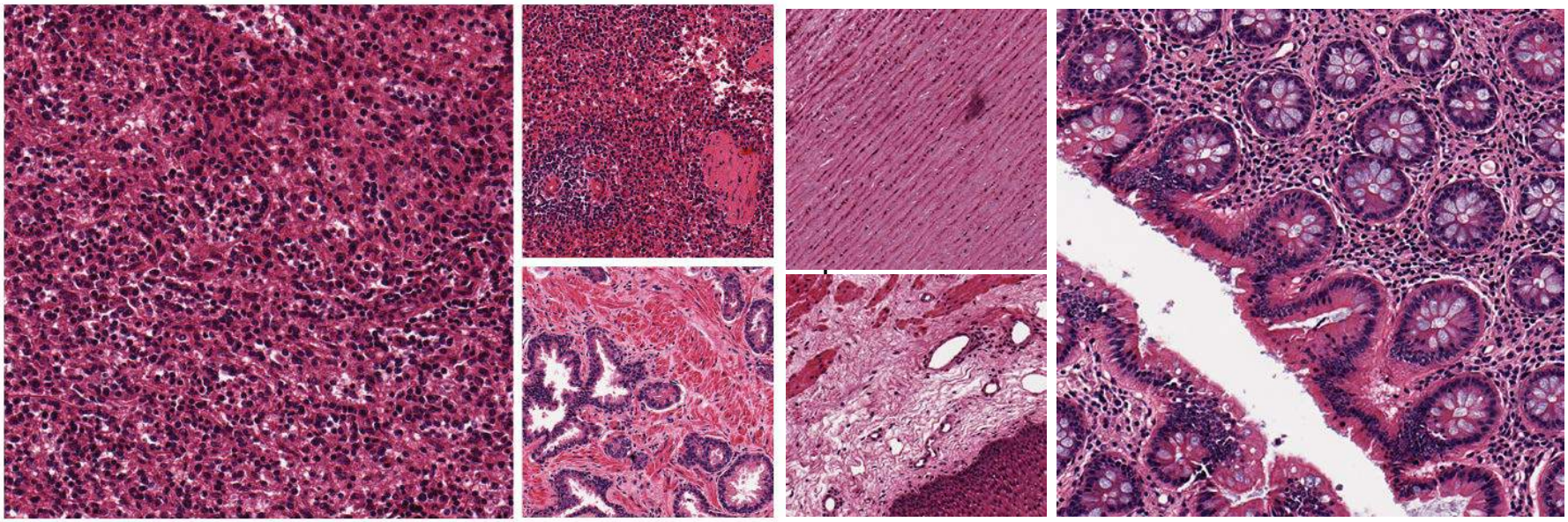
Jordan Ash
Daniel Munro
Greg Darnell

Princeton University

SAGES
6/9/2017



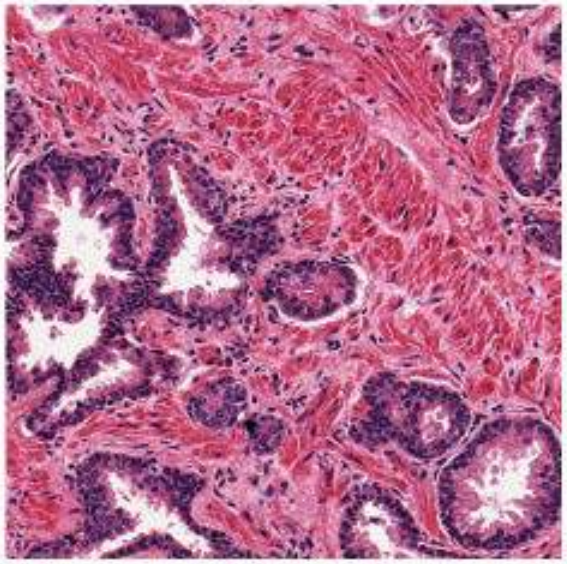
Histological images: a picture's worth a thousand quantitative traits



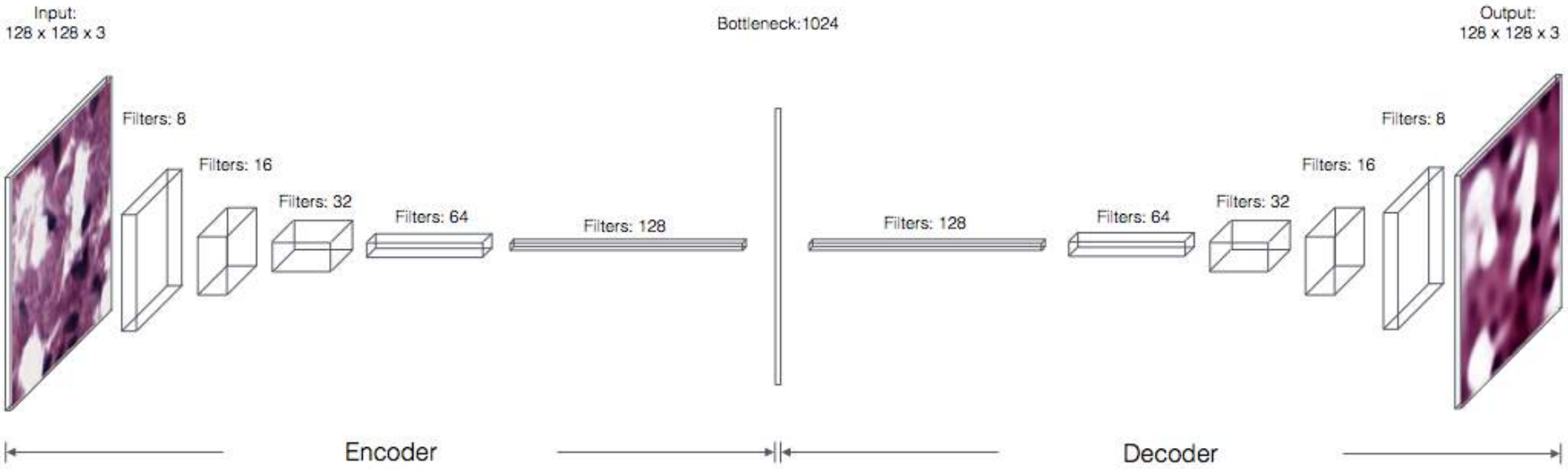
- Histological images used for phenotyping (e.g., cancer diagnosis)
- Features of histological images are associated with:
    - Genotype,
    - Gene expression levels,
    - Cell type,
    - Tissue organization

Quantitative traits from images

- To analyze images, need to characterize and quantify morphology
- Manual annotation of pathology images is infeasible
- Available image segmentation methods are still fairly naïve
- Here, we use an unsupervised deep learning approach to extract features



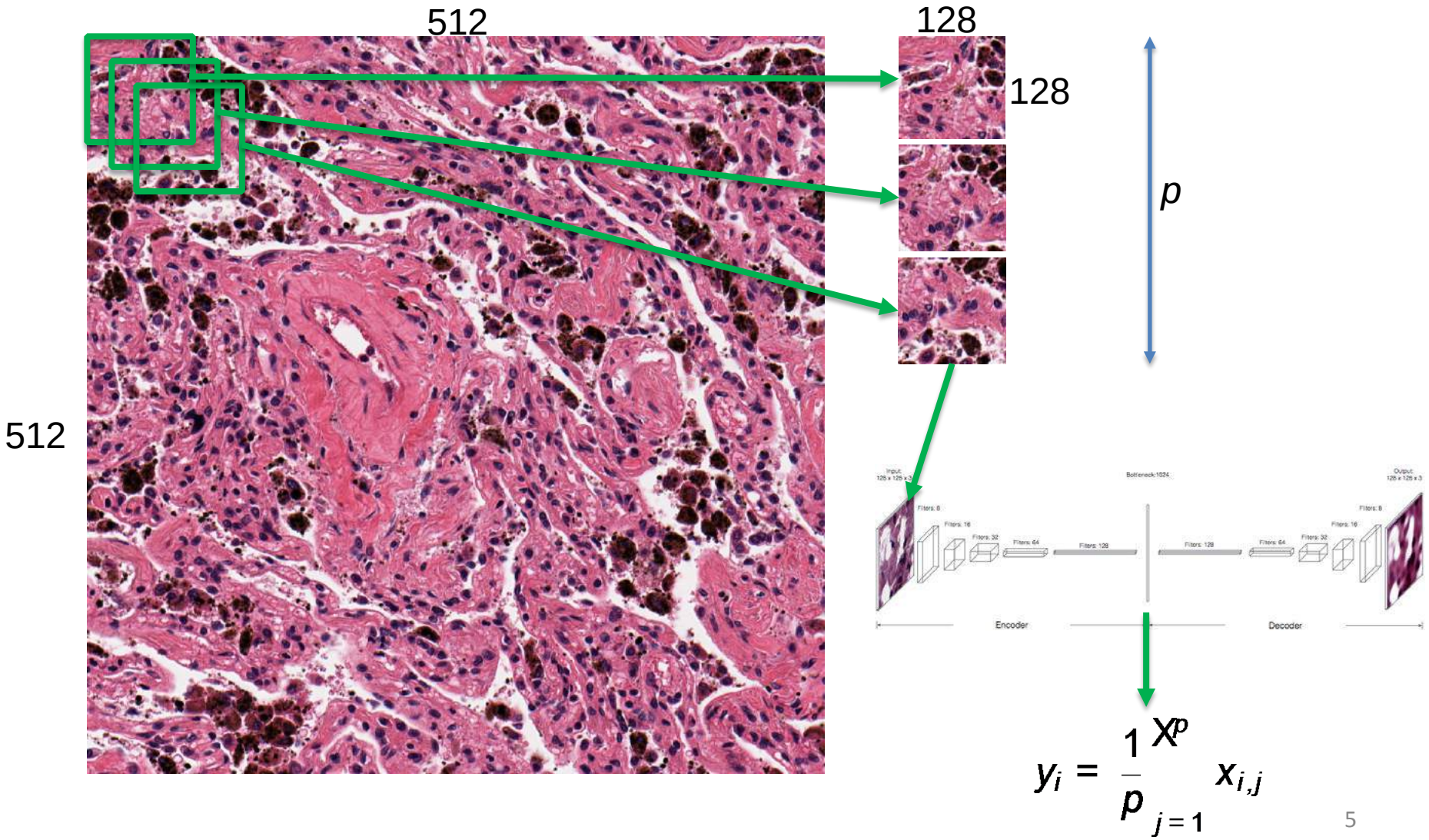
Convolutional autoencoder (CAE)



- Identify 1024 features from CAE
- CAE objective function is perfect reconstruction of image using only 1024 (estimated) features
- Implemented in Keras, interface to TensorFlow

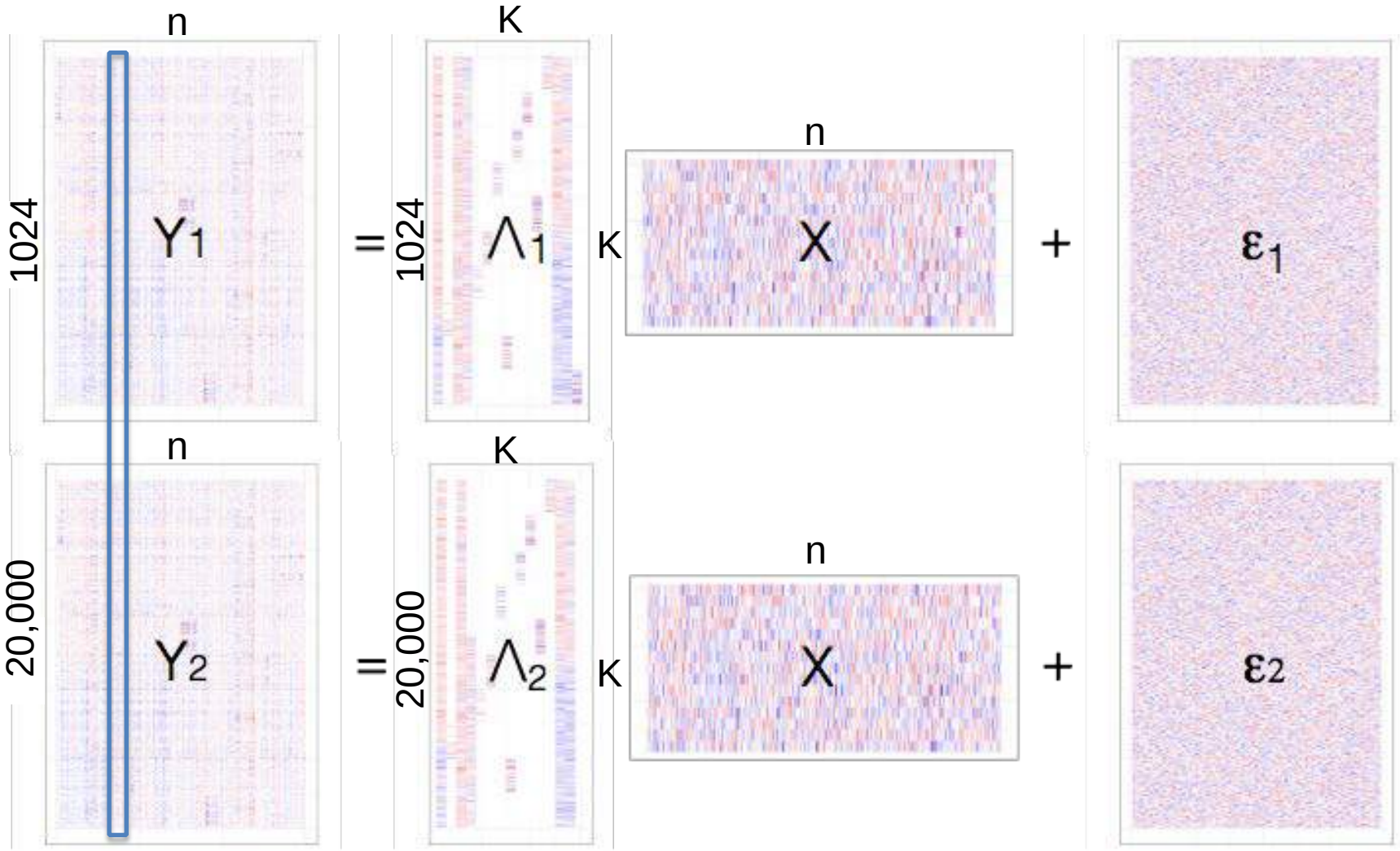
- Question becomes: *what do these image features represent*?

Segmenting each image



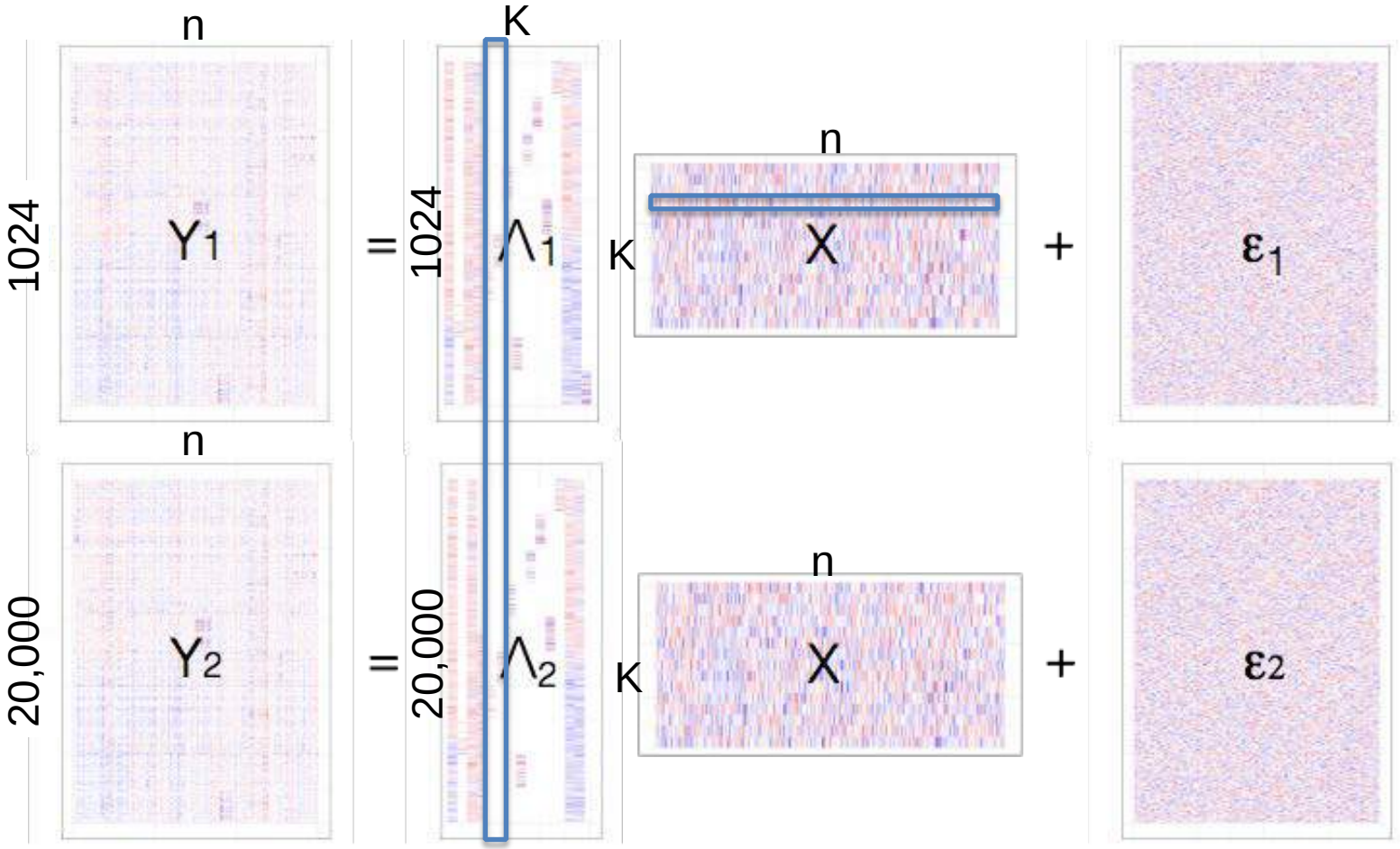
Sparse canonical correlation analysis (CCA)

We now represent images as 1024 quantitative features; correlate with genomic data



[Bach & Jordan, 2005]

Sparse canonical correlation analysis



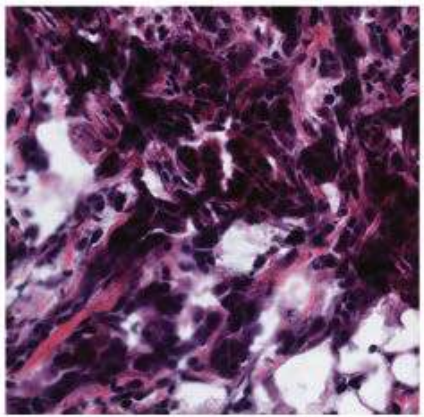
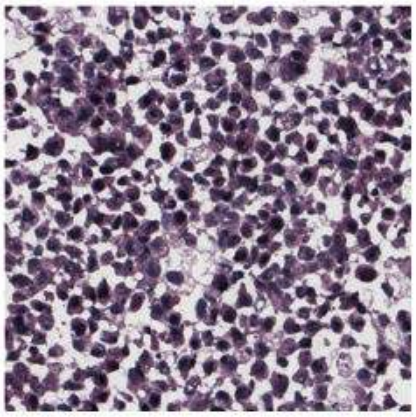
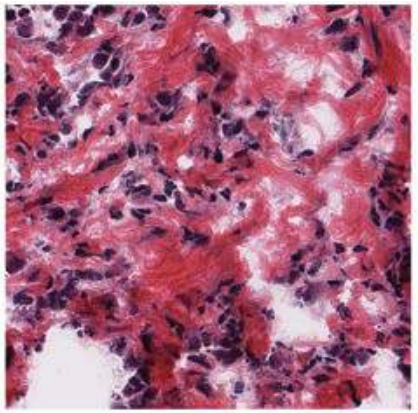
Canonical correlation analysis is a linear projection of two observations into a shared latent subspace that maximizes correlation between observations

Sparsity in loading matrix identifies correlated subsets of genes and image features

[Witten et al. 2009, Zhao et al., 2016]

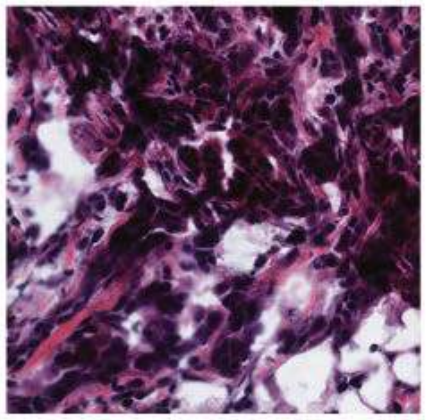
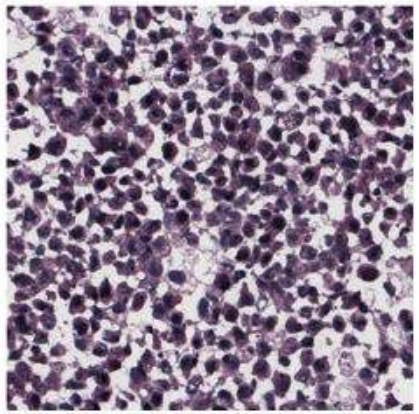
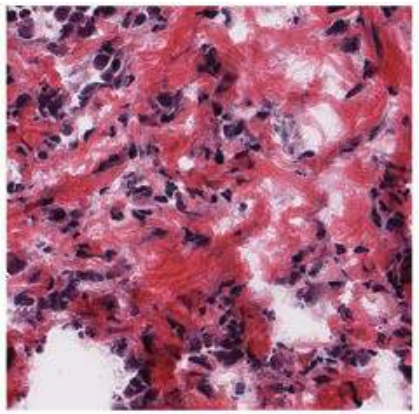
Three applications of ImageCCA

- TCGA: Breast invasive carcinoma study
  - Association of gene expression with tissue features
- TCGA: Brain lower grade glioma
  - Image segmentation
- Genotype-Tissue Expression (GTEx)
  - Identify genetic variants associated with tissue morphology

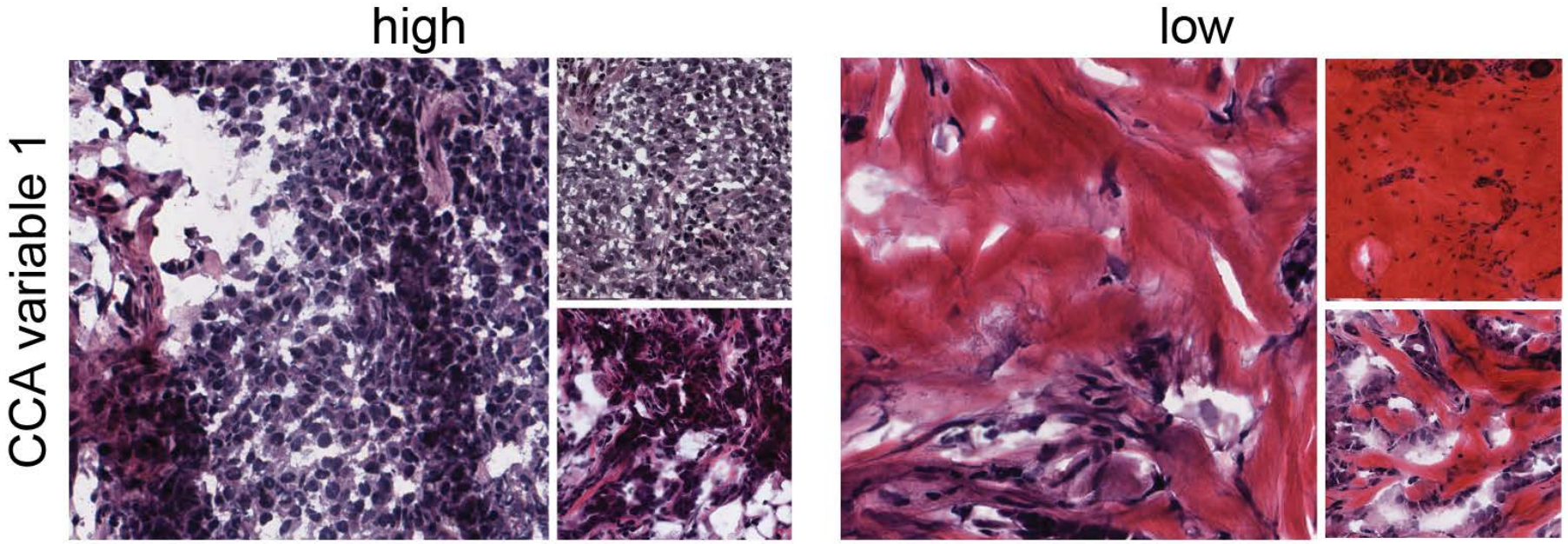




TCGA data: Breast invasive carcinoma study (BRCA)

- 1,541 histological images
  - 1,502 primary tumor samples
  - 7 metastatic tumor samples
  - 32 normal tissue samples
- 1,106 tissue biopsy samples
  - TPM values for 20,501 genes from RNA-seq
- 1,073 breast cancer patients; labels are sample type

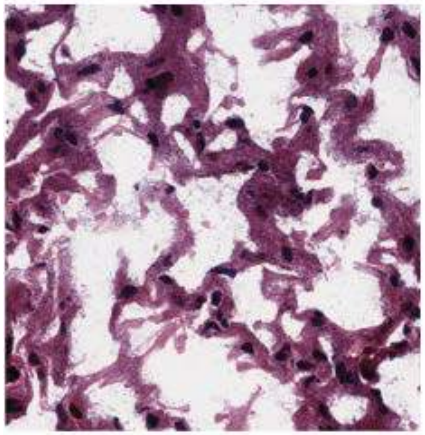
BRCA: component captures extracellular matrix

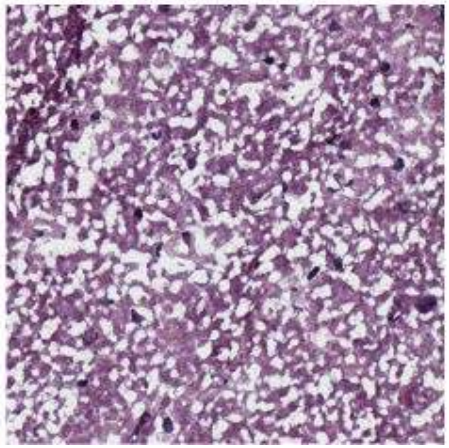


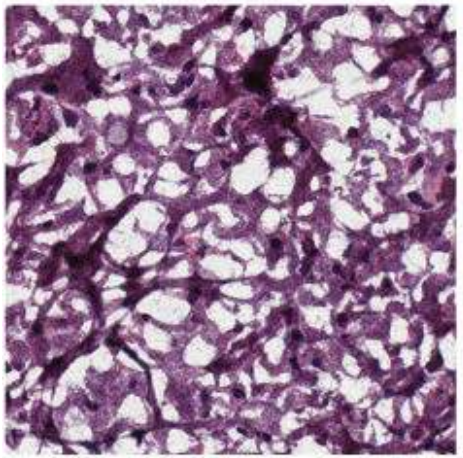
	top GO terms	p
BP	cell adhesion	3e-6
	biological adhesion	3e-6
CC	proteinaceous extracellular matrix	3e-8
	extracellular matrix	2e-7
MF	ion channel binding	1e-3
	collagen binding	2e-3

TCGA data: Brain Lower Grade Glioma (LGG)

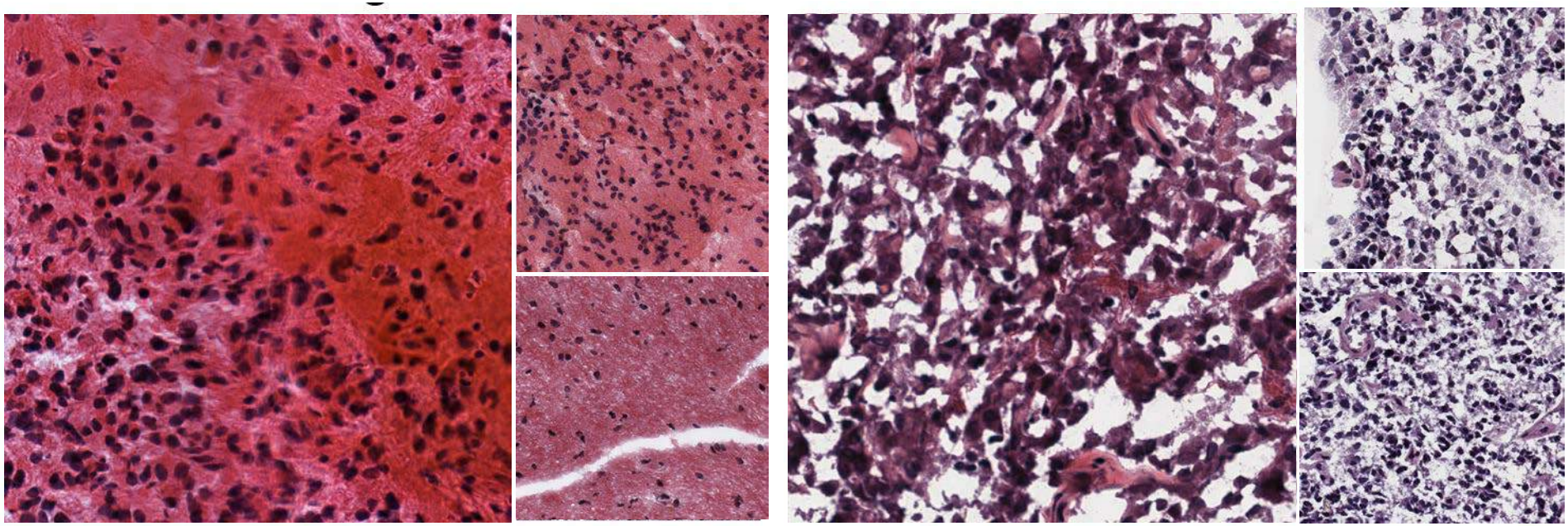
- 484 histological images
    - 471 primary tumor samples
    - 13 recurrent tumor samples
- 401 tissue biopsy samples
    - TPM values for 20,501 genes from RNA-seq
- 392 lower grade glioma patients; labels are sample type





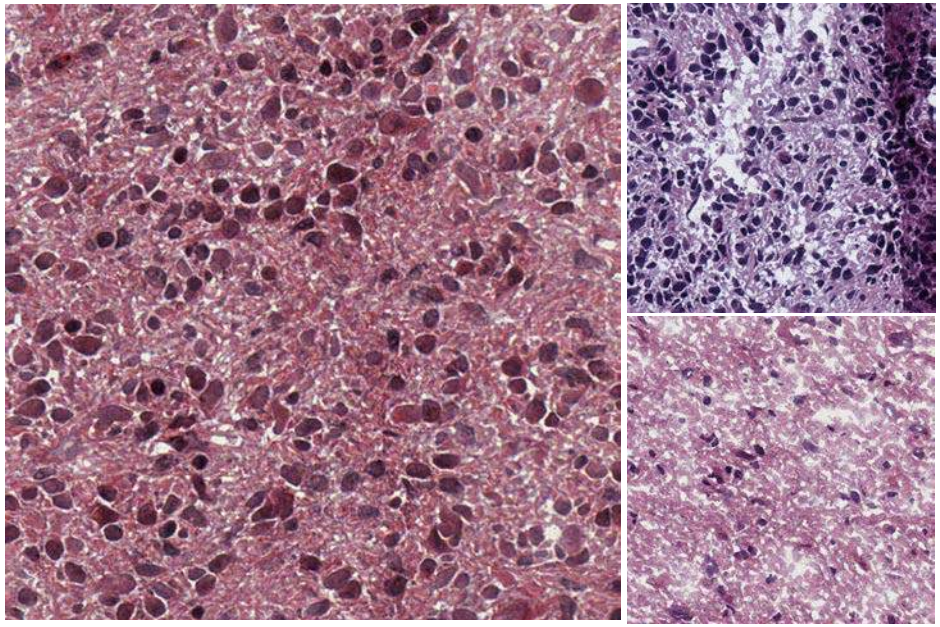


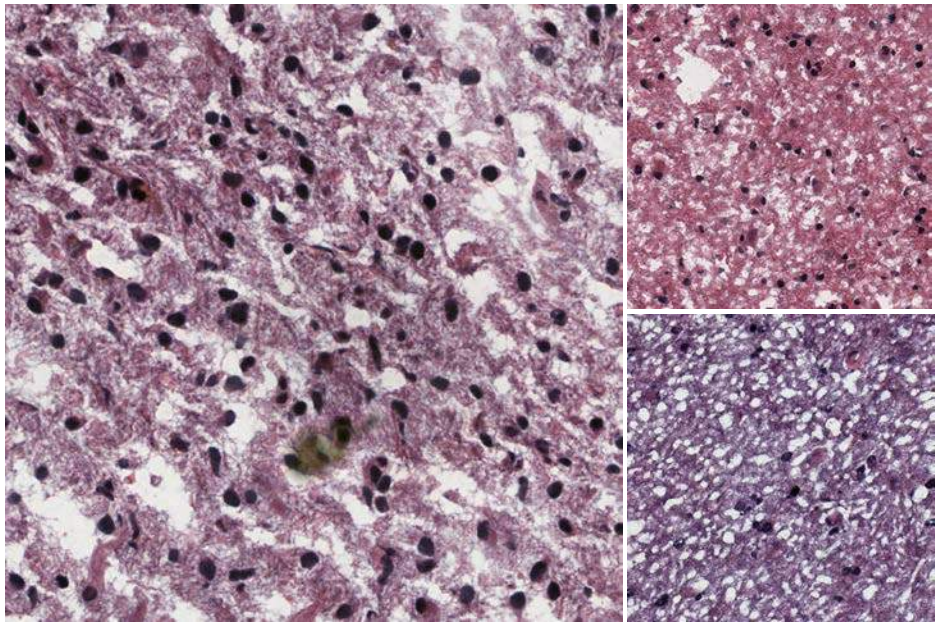
LGG: component represents synaptic structure



Top GO terms	
Synaptic transmission	1.3e-23
Synaptic signaling	1.3e-23
Trans-synaptic signaling	1.3e-23
Cell-cell signaling	5.6e-18

LGG: component represents proportion of blood in brain tissues



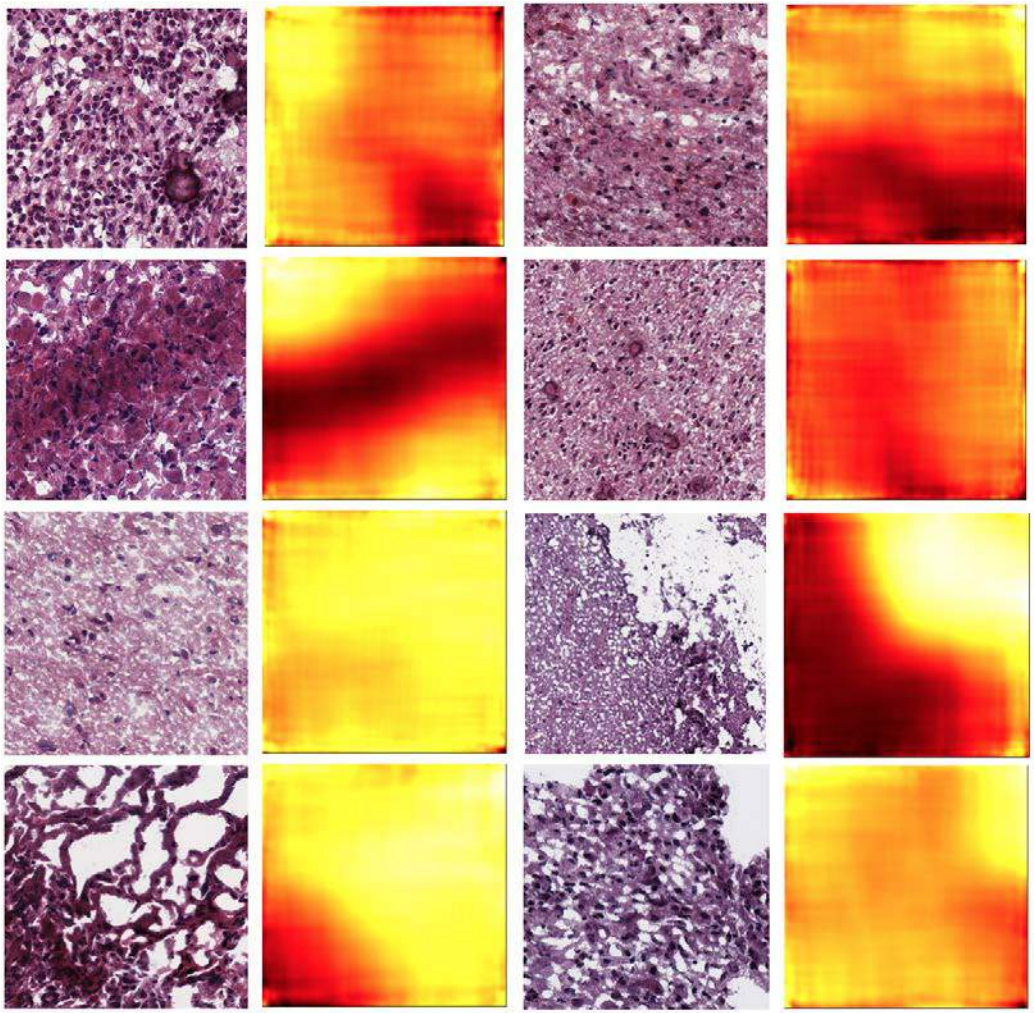


Component genes mainly expressed only in whole blood

Top GO terms	
Immune response	3.9e-29
Immune system process	1.9e-27
Defense response	2.0e-21
Regulation of immune system process	1.1e-20

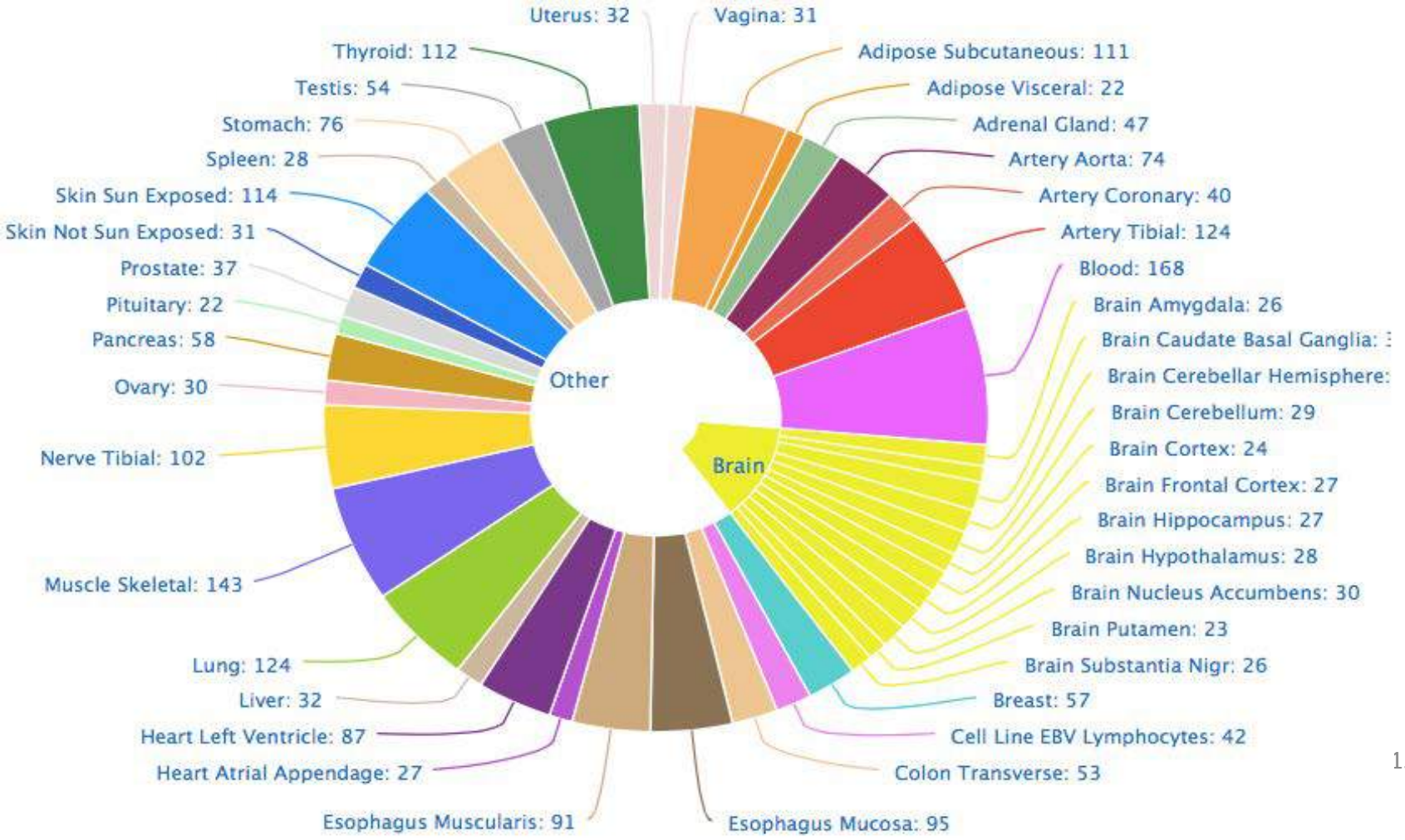
LGG: Identifying cancer in an image

- slide classifier over each 512x512 image; probability of cancer in each 128x128 window
- create a heatmap of these probabilities that highlights regions of the image that the network predicts are cancerous
- darker colors indicate higher probability of cancer



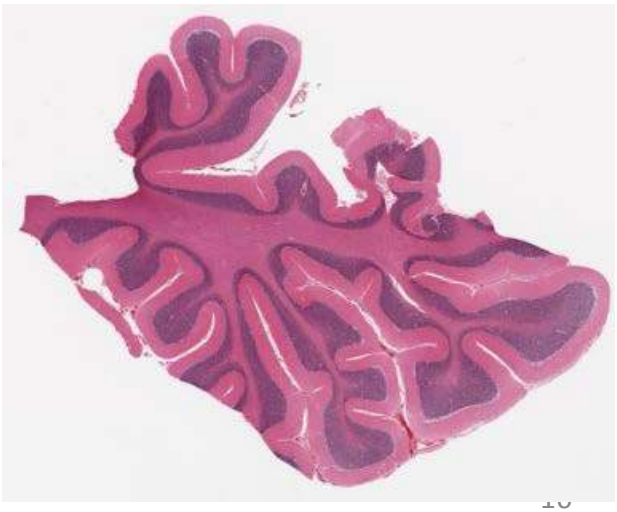
Genotype-Tissue Expression (GTEx) Consortium v6 Data

552 individuals 7,310 samples 4,605 males 2,705 females

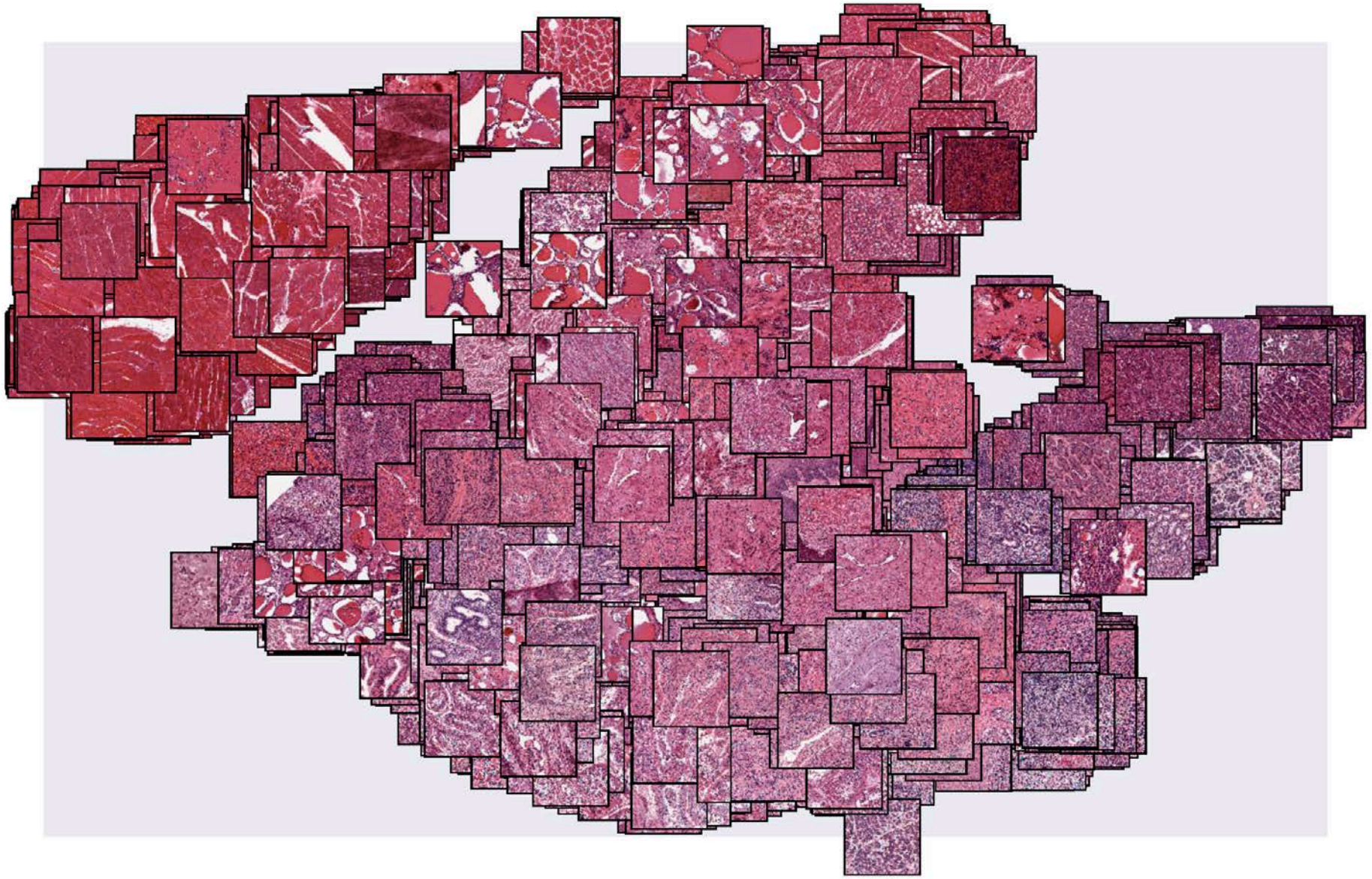


GTEx data: histological images

- 2,221 histological images
    - 29 different tissue types
- 2,221 tissue biopsy samples
    - TPM values for 18,659 genes from RNA-seq
- 499 participants; labels are tissue type

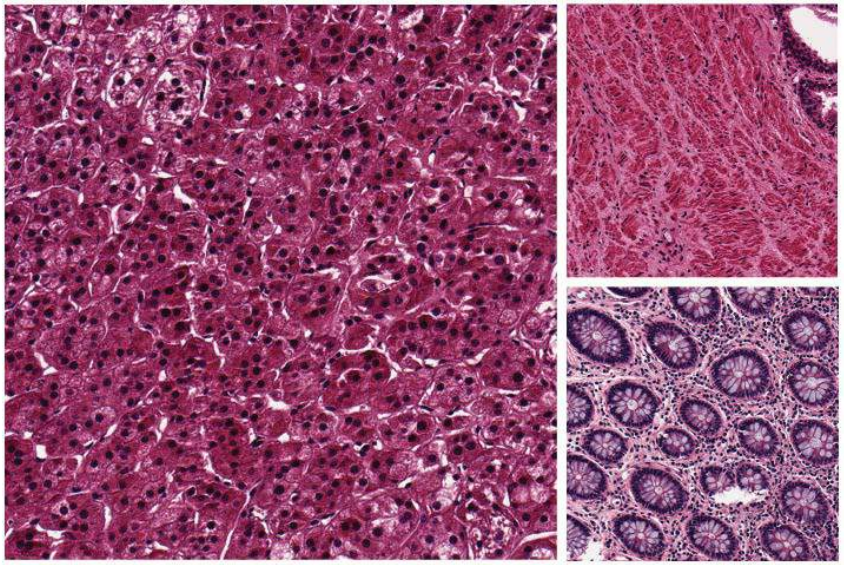


GTEx histological images: t-SNE

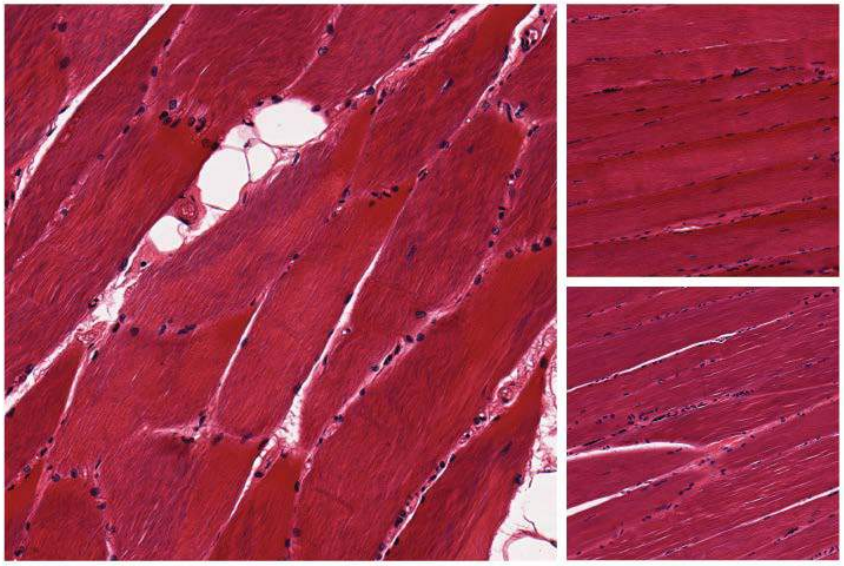


GTEx: component identifies muscle tissues

high

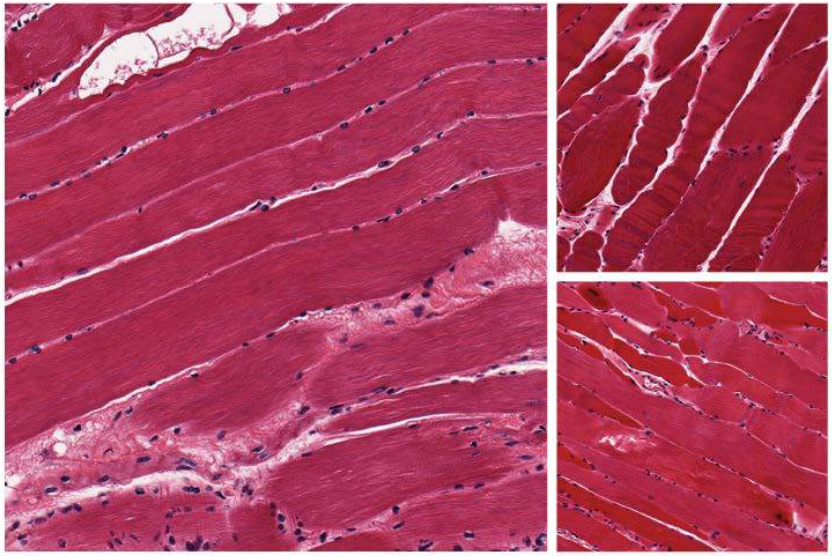


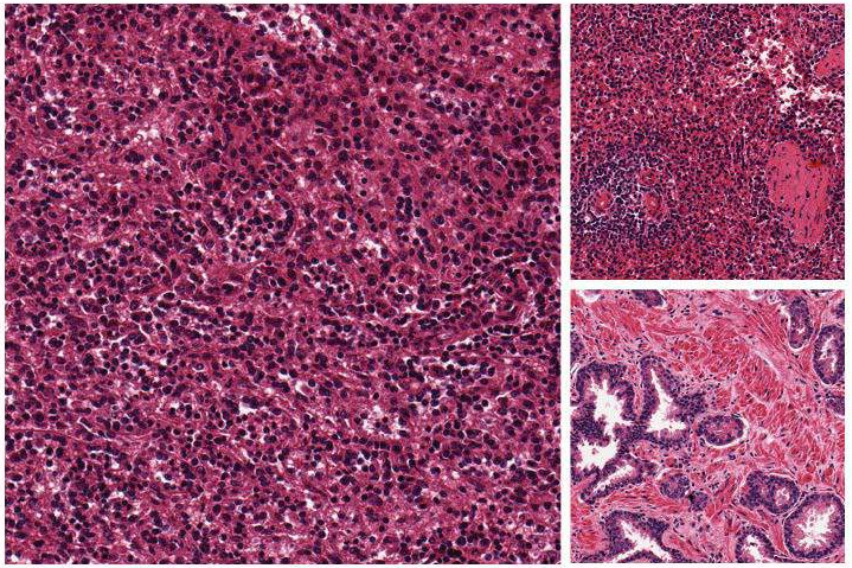
low



	top GO terms	p
BP	muscle system process	< 1e-30
	muscle contraction	< 1e-30
CC	contractile fiber	< 1e-30
	myofibril	6e-30
MF	actin binding	2e-11
	structural constituent of muscle	5e-11

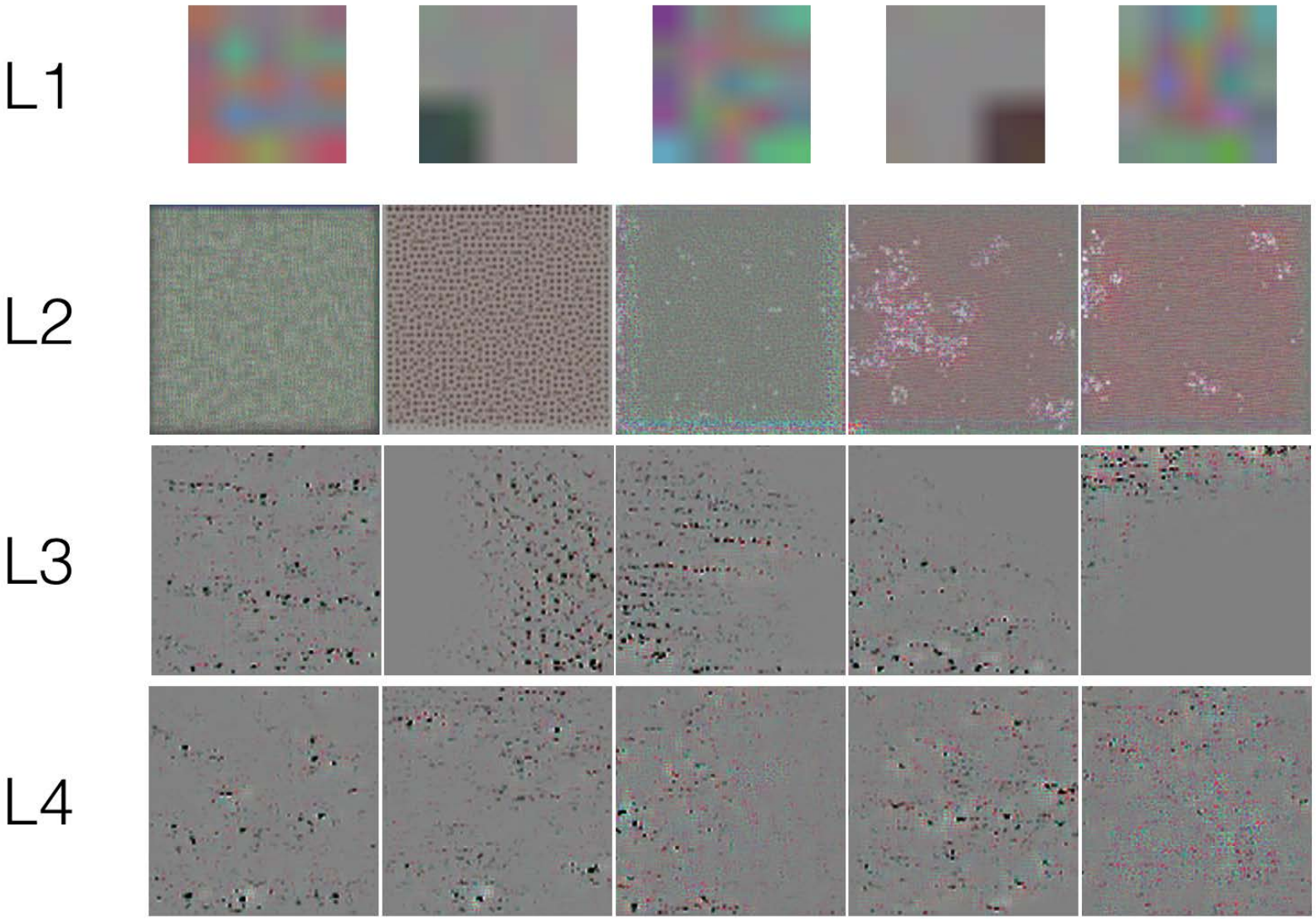
GTEx: component identifies neuronal tissues



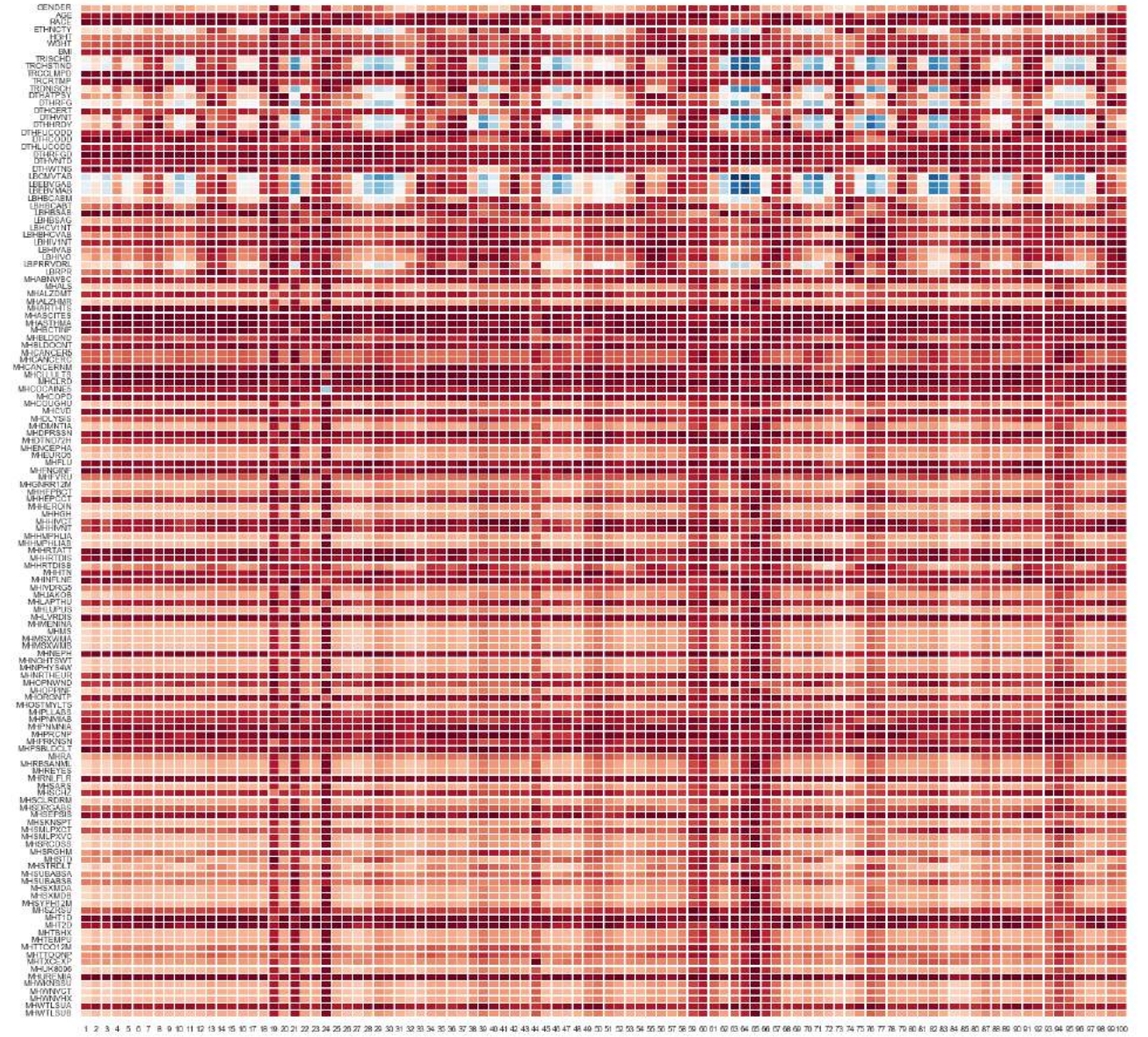


	top GO terms	p
BP	synaptic transmission, cholinergic	2e-5
	muscle organ development	9e-5
CC	acetylcholine-gated channel complex	6e-7
	myofibril	2e-5
MF	acetylcholine-activated cation-selective channel activity	5e-7
	acetylcholine binding	9e-7

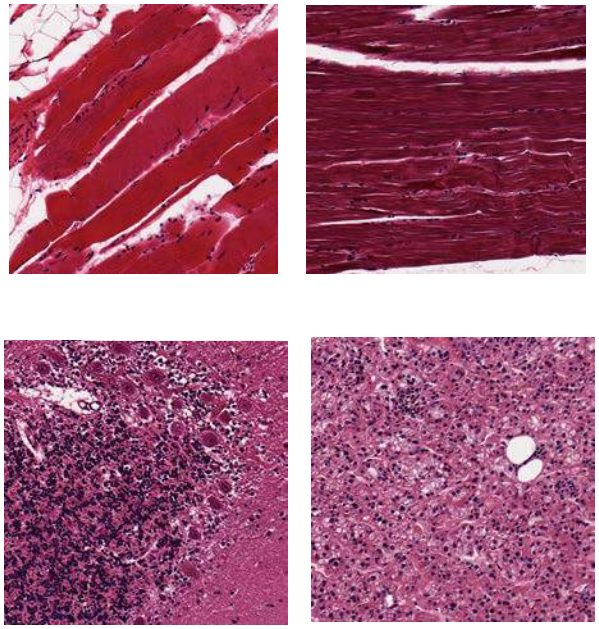
Convolutional filters estimated



CCA components correlate with covariates



Genetic associations with image features



rs11102981 synaptophysin-like 2 (SYPL2)

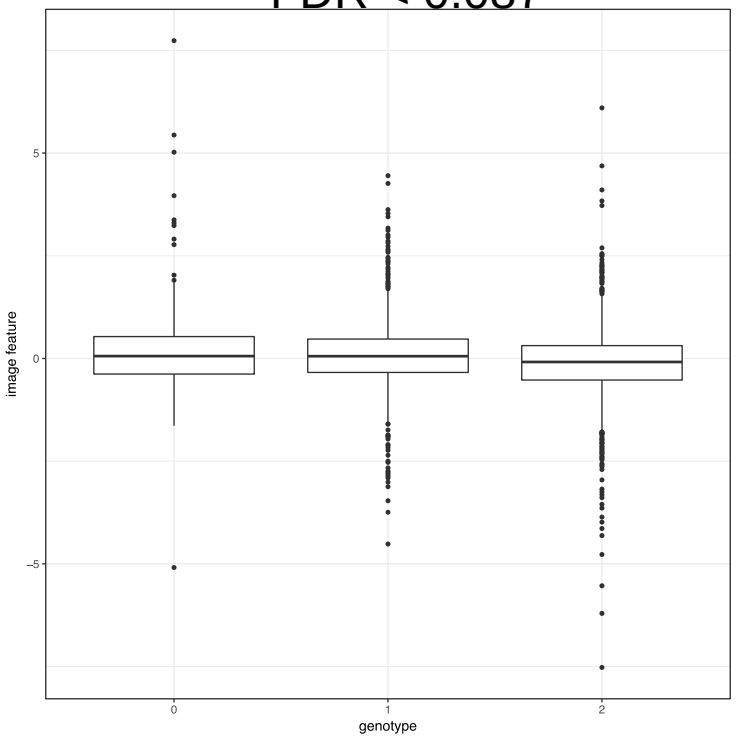
FDR < 0.087

rs68178377 protein phosphatase 6 (PPP6R2)

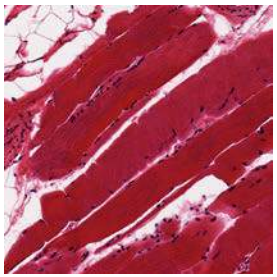
FDR < 0.10

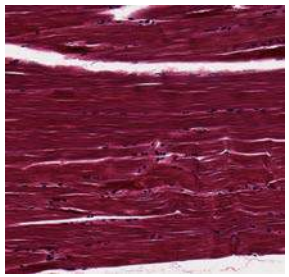
Genetic association with image feature

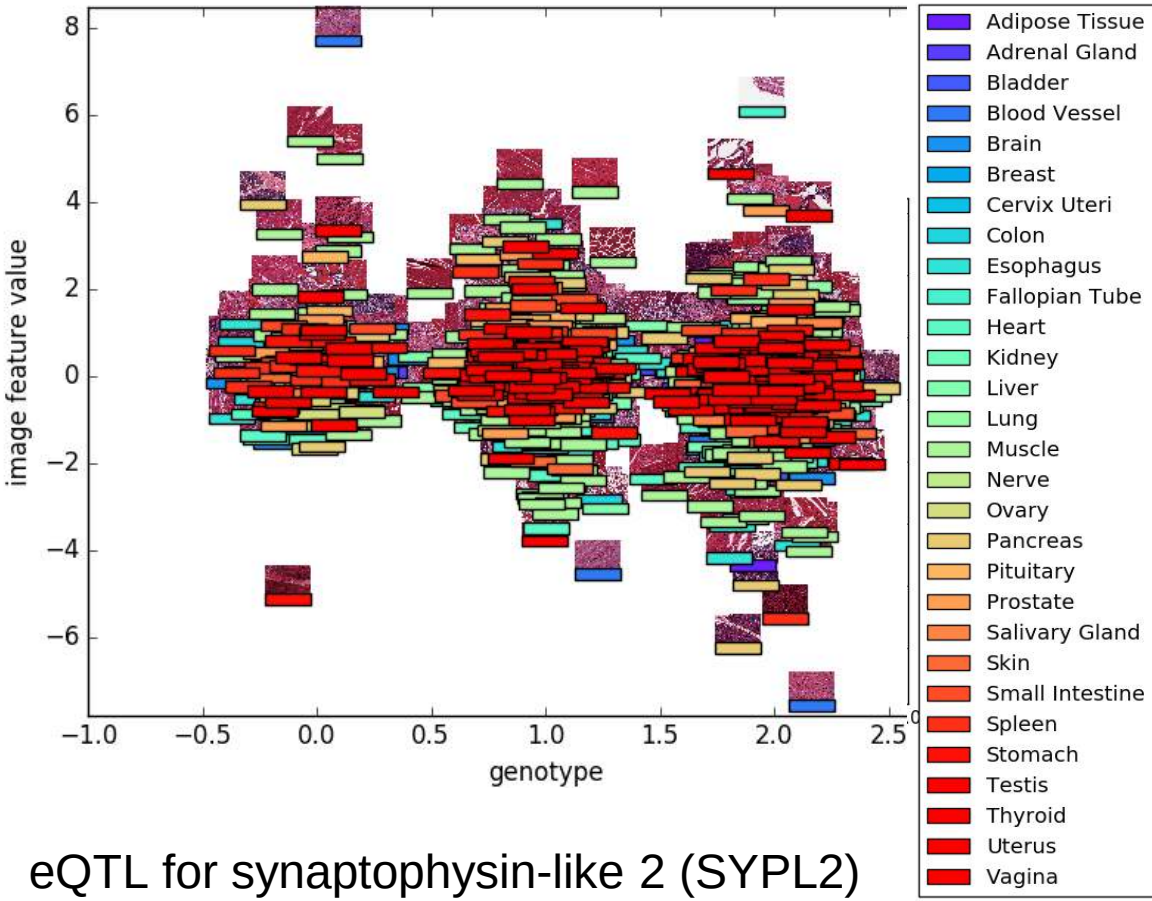
FDR < 0.087



rs11102981

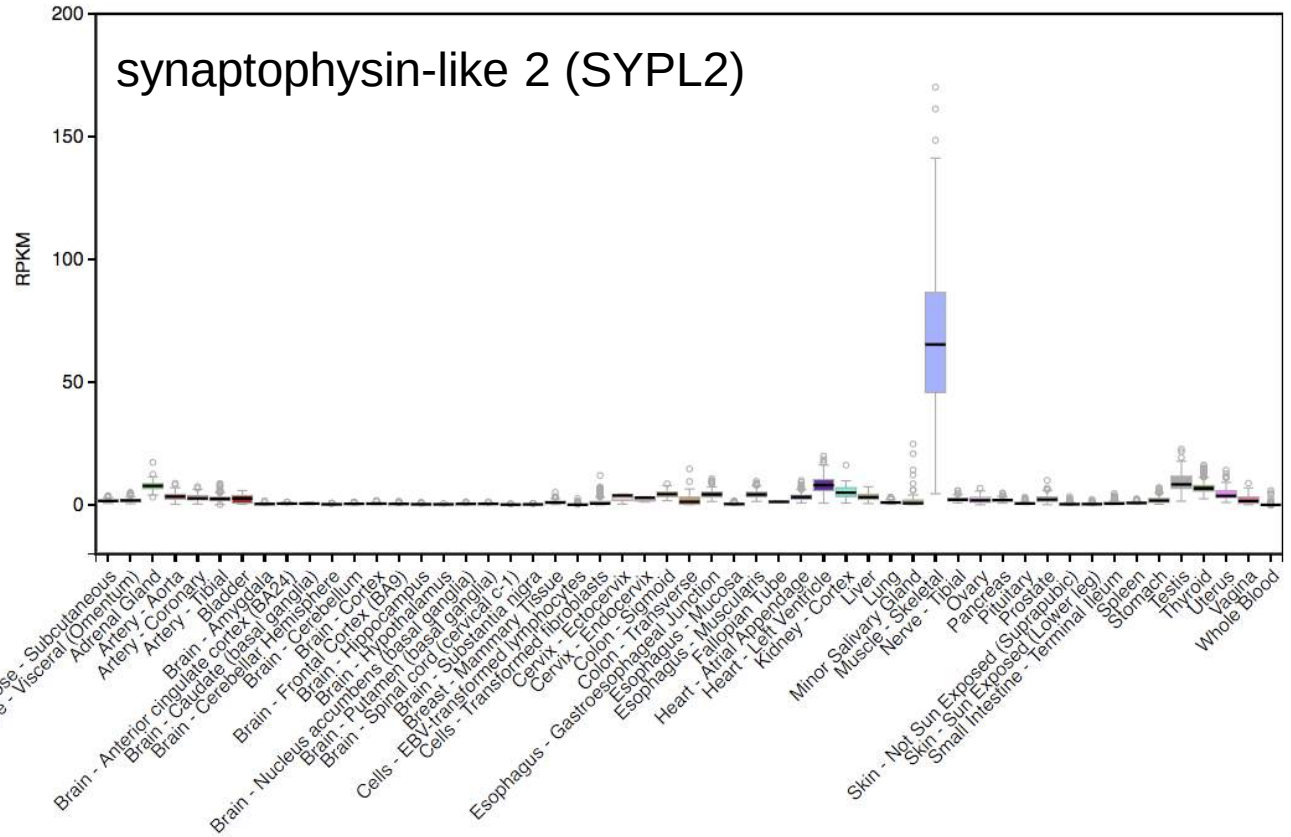


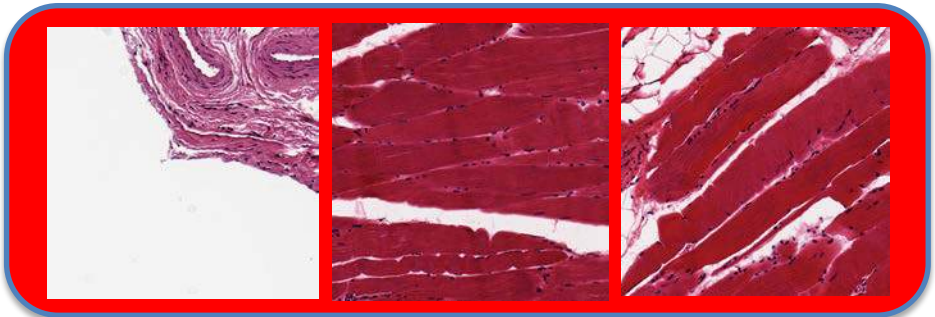




eQTL for synaptophysin-like 2 (SYPL2)
Involved in communication between
T-tubular and junctional sarcoplasmic reticulum membranes

Image phenotype associated with muscle tissue morphology





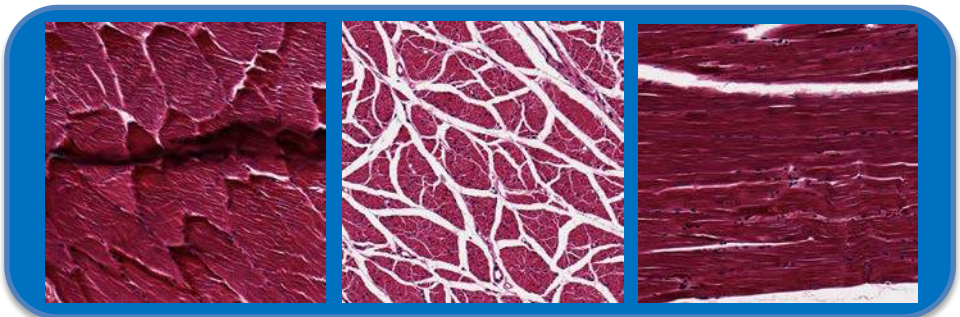
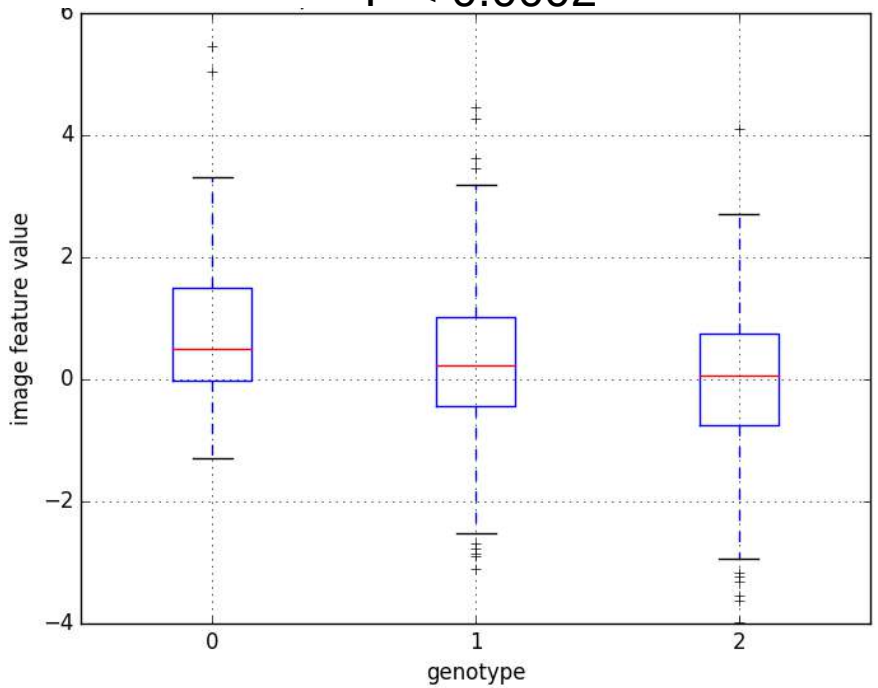
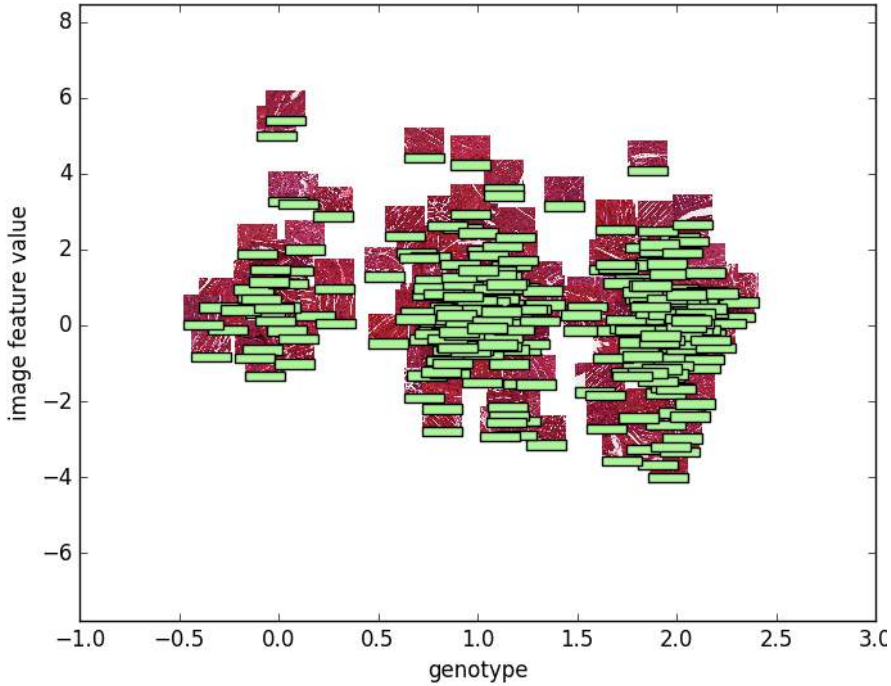


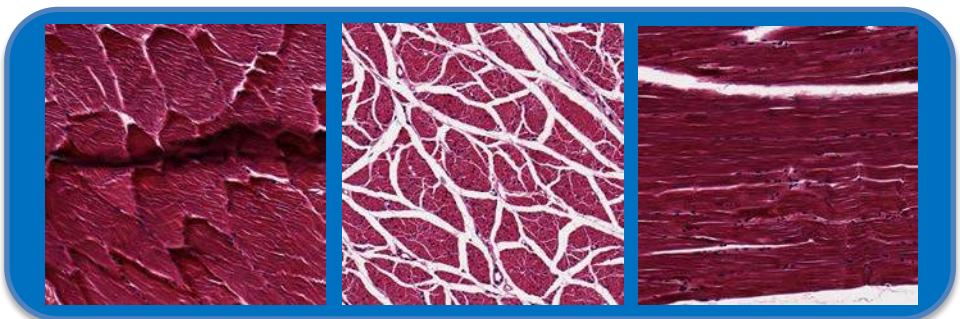
Image phenotype associated with muscle tissue morphology

P < 0.0002



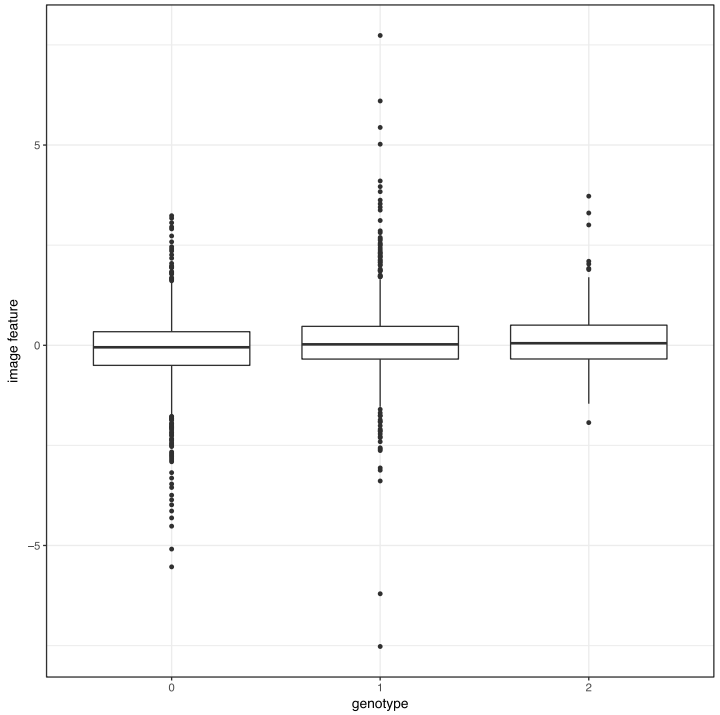
rs11102981

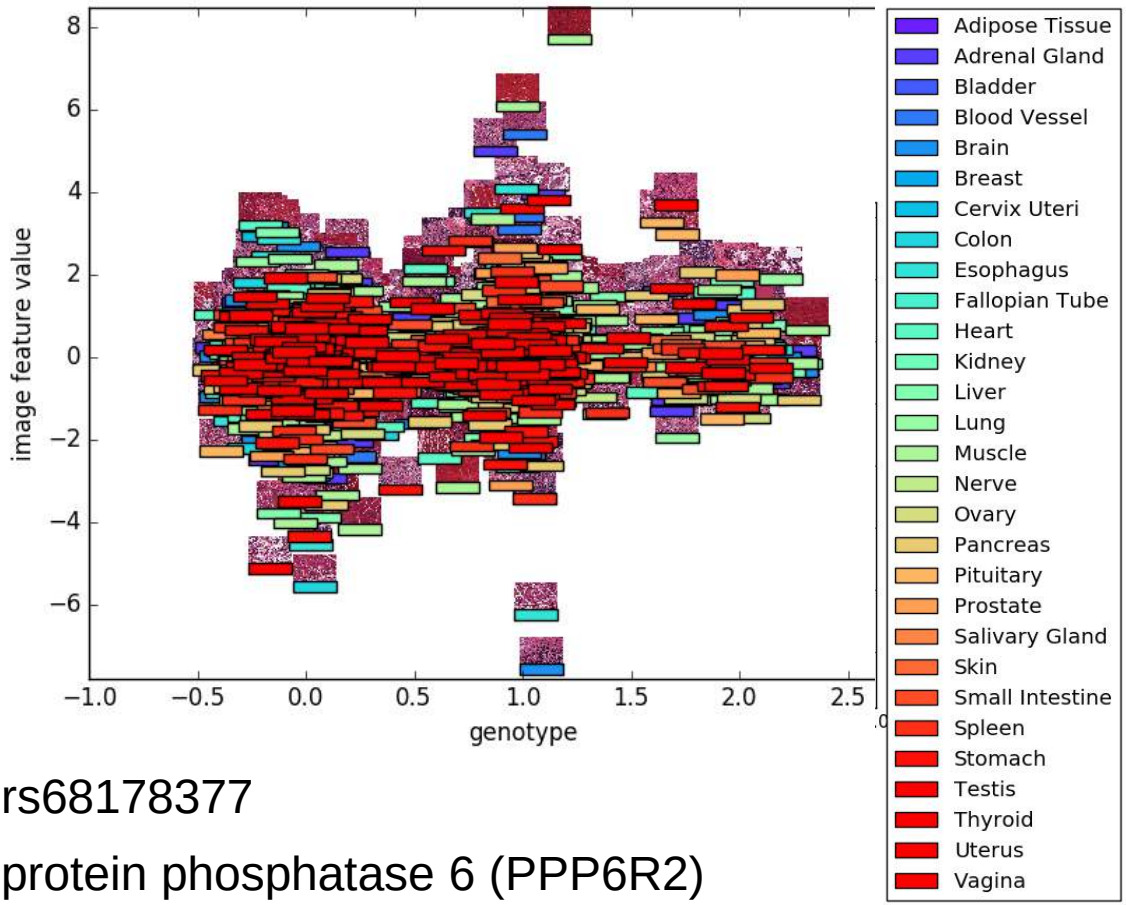


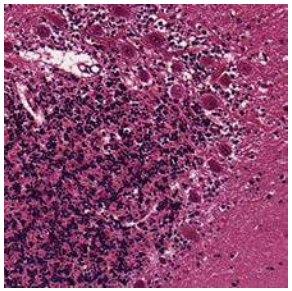


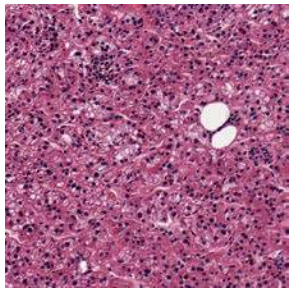
Genetic association with image feature

FDR < 0.10





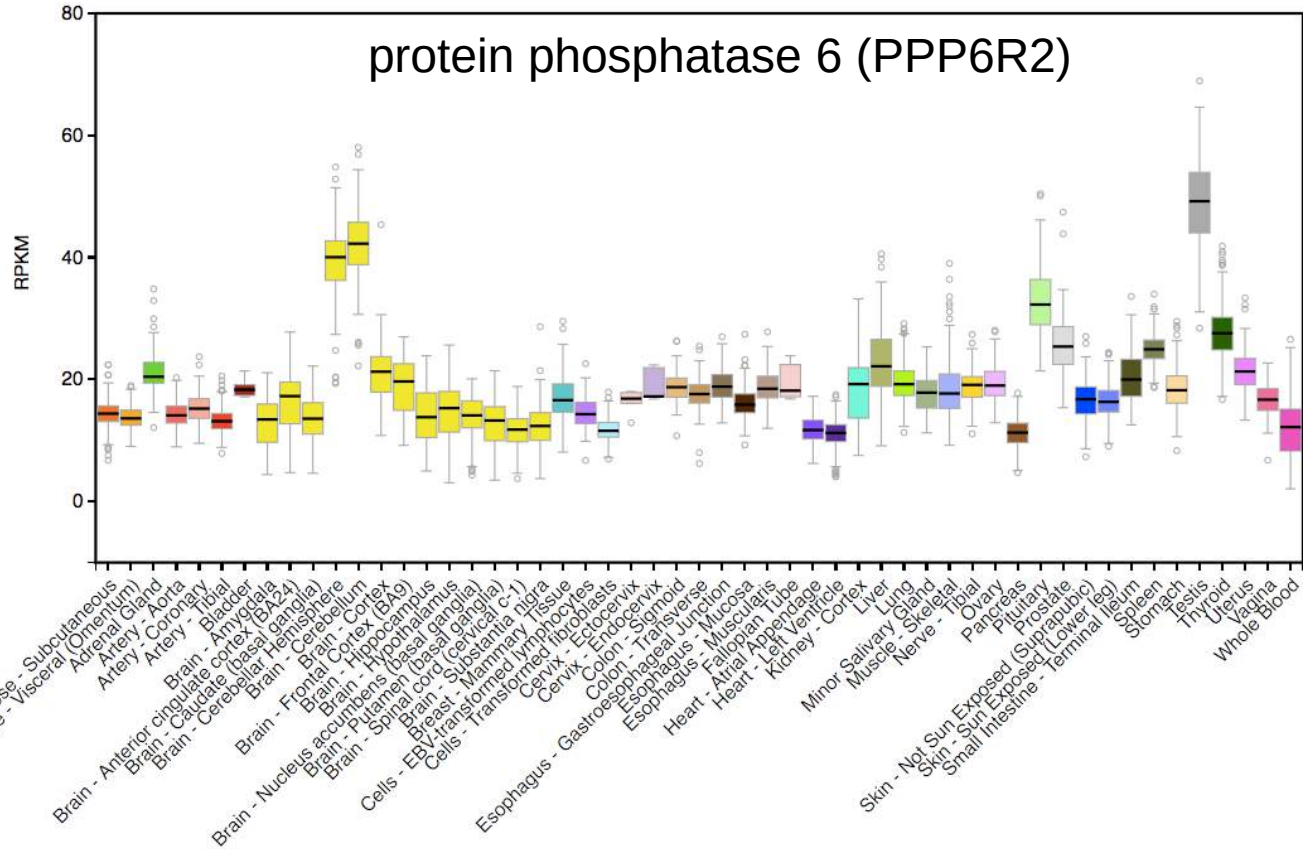


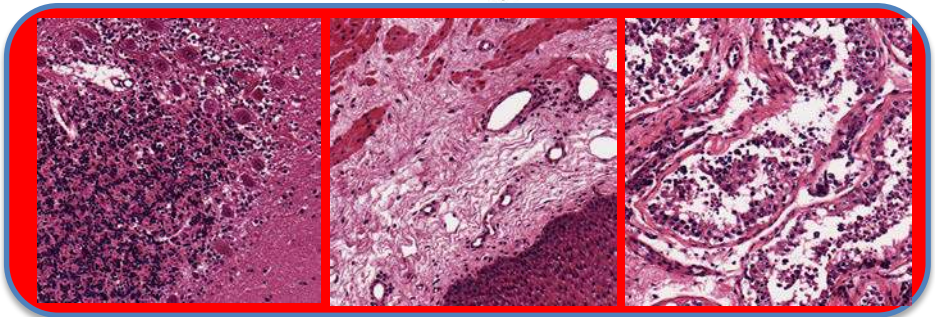


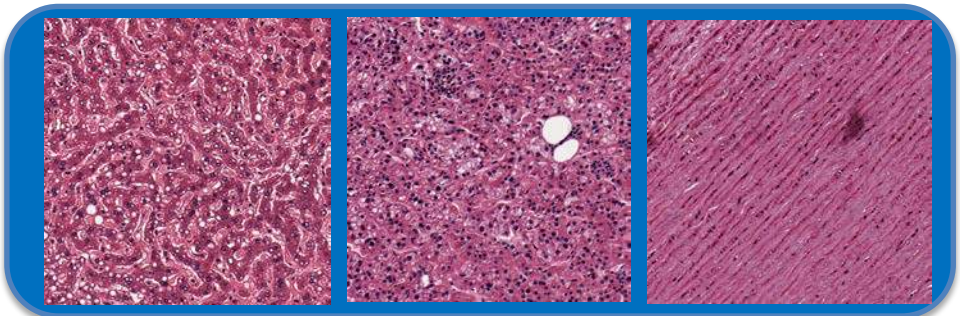
rs68178377

protein phosphatase 6 (PPP6R2)

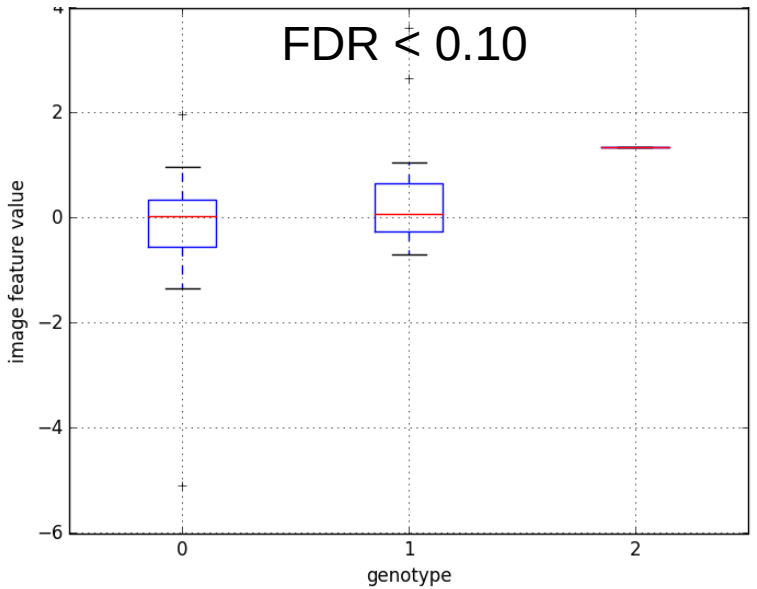
Image phenotype associated with brain tissue morphology

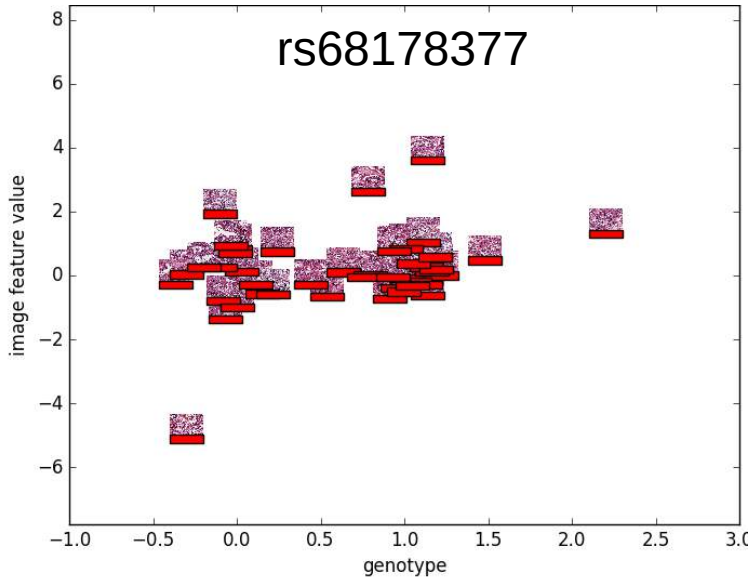






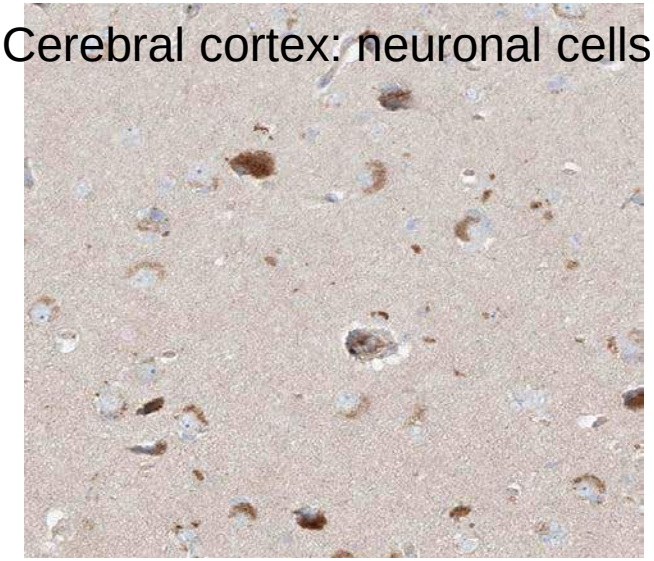
Association of image phenotype driving brain and testis morphology



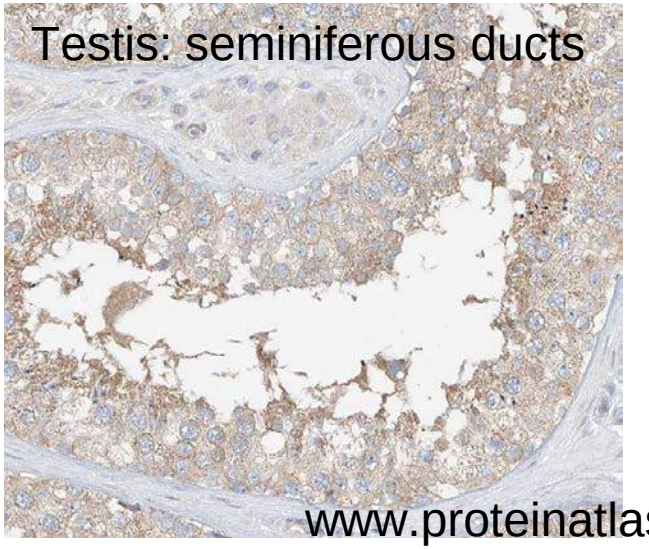


protein phosphatase 6 (PPP6R2)

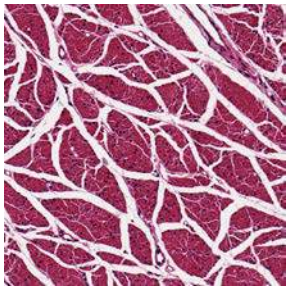
Cerebral cortex: neuronal cells



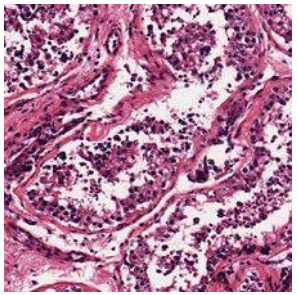
Testis: seminiferous ducts



www.proteinatlas.org



Conclusions



- Automatically extracted features from histology images
- Correlated image features with high dimensional gene expression
- Components captured tissue type, cell type heterogeneity, and morphological features of data
- Identified genetic variants associated with specific image features in muscle and brain/testis

Acknowledgements

Collaborators

- Sayan Mukherjee (Duke)
- GTEx Consortium
  - LDACC: Kristin Ardlie
  - Pathology imaging lab:
    - Phil Branton

Princeton University

- Jordan Ash
- Daniel Munro
- Greg Darnell

Funding

- NIH R01 MH101822
- Sloan Faculty Fellowship

Visualizing the GTEx decoder

- Generated random 1024 vectors on the manifold of images
- pushed through GTEx decoder to generate an image