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

\[ y_i = \frac{1}{p} \sum_{j=1}^{p} x_{i,j} \]
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**

<table>
<thead>
<tr>
<th>CCA variable 1</th>
<th>high</th>
<th>low</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>top GO terms</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>BP cell adhesion</td>
<td>3e-6</td>
</tr>
<tr>
<td>BP biological adhesion</td>
<td>3e-6</td>
</tr>
<tr>
<td>CC proteinaceous extracellular matrix</td>
<td>3e-8</td>
</tr>
<tr>
<td>CC matrix</td>
<td>2e-7</td>
</tr>
<tr>
<td>MF ion channel binding</td>
<td>1e-3</td>
</tr>
<tr>
<td>MF collagen binding</td>
<td>2e-3</td>
</tr>
</tbody>
</table>
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: $1.3 \times 10^{-23}$
- Synaptic signaling: $1.3 \times 10^{-23}$
- Trans-synaptic signaling: $1.3 \times 10^{-23}$
- Cell-cell signaling: $5.6 \times 10^{-18}$
LGG: component represents proportion of blood in brain tissues

Component genes mainly expressed only in whole blood

**Top GO terms**
- Immune response: $3.9 \times 10^{-29}$
- Immune system process: $1.9 \times 10^{-27}$
- Defense response: $2.0 \times 10^{-21}$
- Regulation of immune system process: $1.1 \times 10^{-20}$
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

**top GO terms**

<table>
<thead>
<tr>
<th>Category</th>
<th>Term</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>BP</td>
<td>muscle system process</td>
<td>&lt; 1e-30</td>
</tr>
<tr>
<td></td>
<td>muscle contraction</td>
<td>&lt; 1e-30</td>
</tr>
<tr>
<td>CC</td>
<td>contractile fiber</td>
<td>&lt; 1e-30</td>
</tr>
<tr>
<td></td>
<td>myofibril</td>
<td>6e-30</td>
</tr>
<tr>
<td>MF</td>
<td>actin binding</td>
<td>2e-11</td>
</tr>
<tr>
<td></td>
<td>structural constituent of muscle</td>
<td>5e-11</td>
</tr>
</tbody>
</table>
GTEx: component identifies neuronal tissues

<table>
<thead>
<tr>
<th>top GO terms</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>BP synaptic transmission, cholinergic muscle organ</td>
<td>2e-5</td>
</tr>
<tr>
<td>development</td>
<td>9e-5</td>
</tr>
<tr>
<td>CC acetylcholine-gated channel complex</td>
<td>6e-7</td>
</tr>
<tr>
<td>myofibril</td>
<td>2e-5</td>
</tr>
<tr>
<td>MF acetylcholine-activated cation-selective channel</td>
<td>5e-7</td>
</tr>
<tr>
<td>activity acetylcholine binding</td>
<td>9e-7</td>
</tr>
</tbody>
</table>
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

rs11102981

FDR < 0.087

eQTL for synaptophysin-like 2 (SYPL2)
Involved in communication between T-tubular and junctional sarcoplasmic reticulum membranes
Image phenotype associated with muscle tissue morphology

synaptophysin-like 2 (SYPL2)
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

protein phosphatase 6 (PPP6R2)
Association of image phenotype driving brain and testis morphology

FDR < 0.10

Cerebral cortex: neuronal cells

protein phosphatase 6 (PPP6R2)

Testis: seminiferous ducts

rs68178377
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