Fusion of image and genetic data with convolutional neural networks and AJIVE

Iain Carmichael
University of North Carolina at Chapel Hill
Joint work* with: J.S. Marron, Jan Hannig, Katherine Hoadley and Ben Calhoun

Email: iain@unc.edu
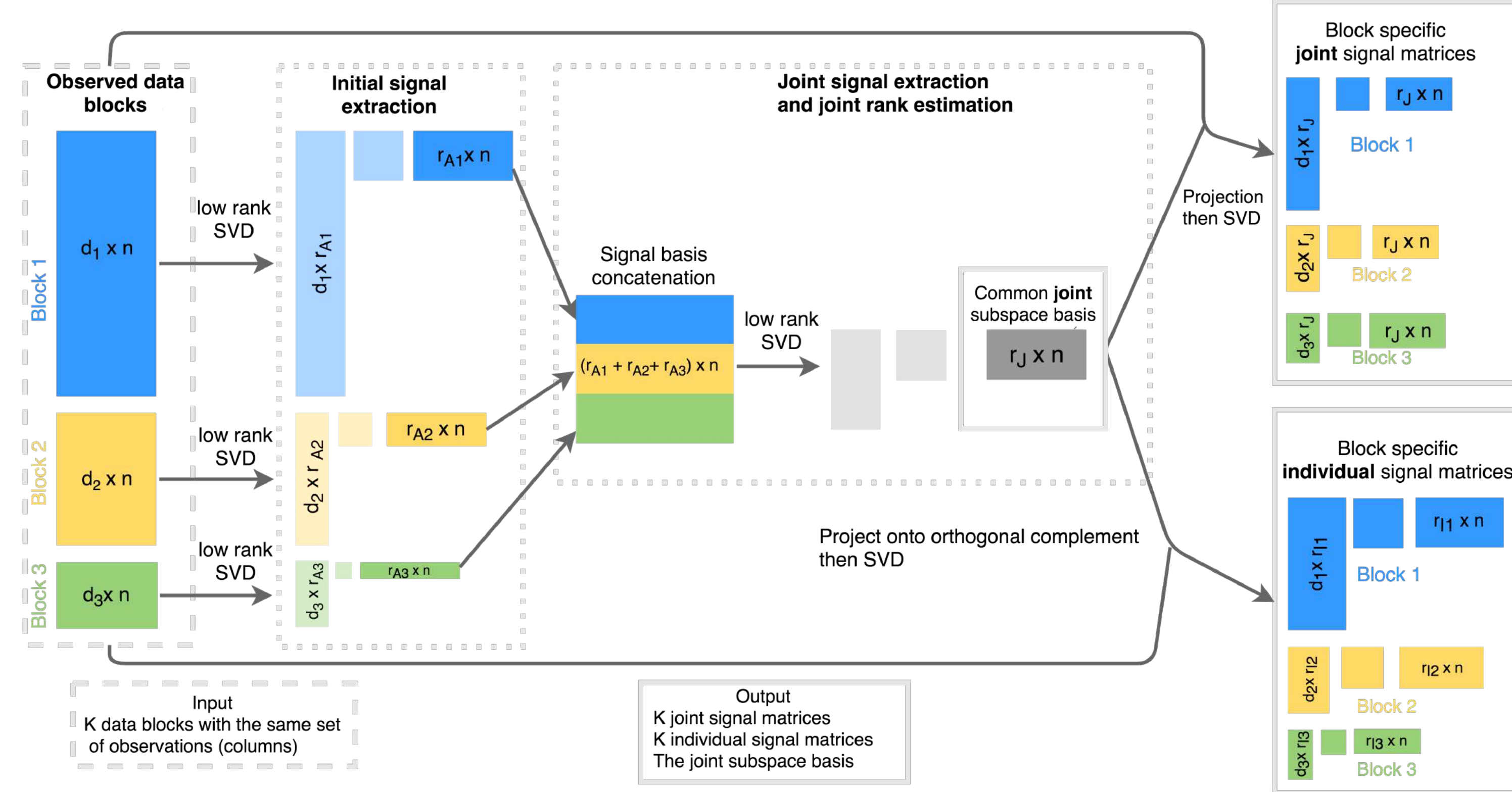Website: idc9.github.io
*(preprint forthcoming)

UNC Statistics & OR

Angle-based joint and individual variation explained (AJIVE)

AJIVE Path Diagram

Multi-block data: fixed set of observations, multiple sets of variables

AJIVE finds joint signals, if any exist, which are common to all data blocks as well as individual signals which are specific to each block, if they exist.

AJIVE matrix decomposition for B data blocks:

$$X_b = J_b + I_b + E_b, \ \ b = 1, \ldots, B$$

Joint space $\mathrm{col}(J_1) = \cdots = \mathrm{col}(J_B)$

Individual spaces $\mathrm{col}(I_b) \perp \mathrm{col}(J_1), \ \ b = 1, \ldots, B$

Statistical inference to segment joint from individual signal using
- Random direction bound
- Wedin bound

Application to medical imaging and breast cancer genetics

Carolina Breast Cancer Study
- 1700 women with breast cancer
- H&E stained tumor biopsy images for each patient
- PAM50 gene expressions for each patient
- Other covariates (e.g. age, pathology report)

Convolutional neural network for image feature extraction
- Patch based representation of large, heterogeneous images
- Subjects are represented via average of patches which contain sufficient tissue
- Pre-trained VGG16

Neural network interpretability methods
- Patch based localization
- New saliency map approaches

AJIVE is used to find connections between histopathology and genetics as well as separate histopathology individual and genetic individual modes of variation
- 7 joint components found

Histopathology image individual

Showing
- Representative subjects and patches from different individual components
- Subjects selected using individual scores
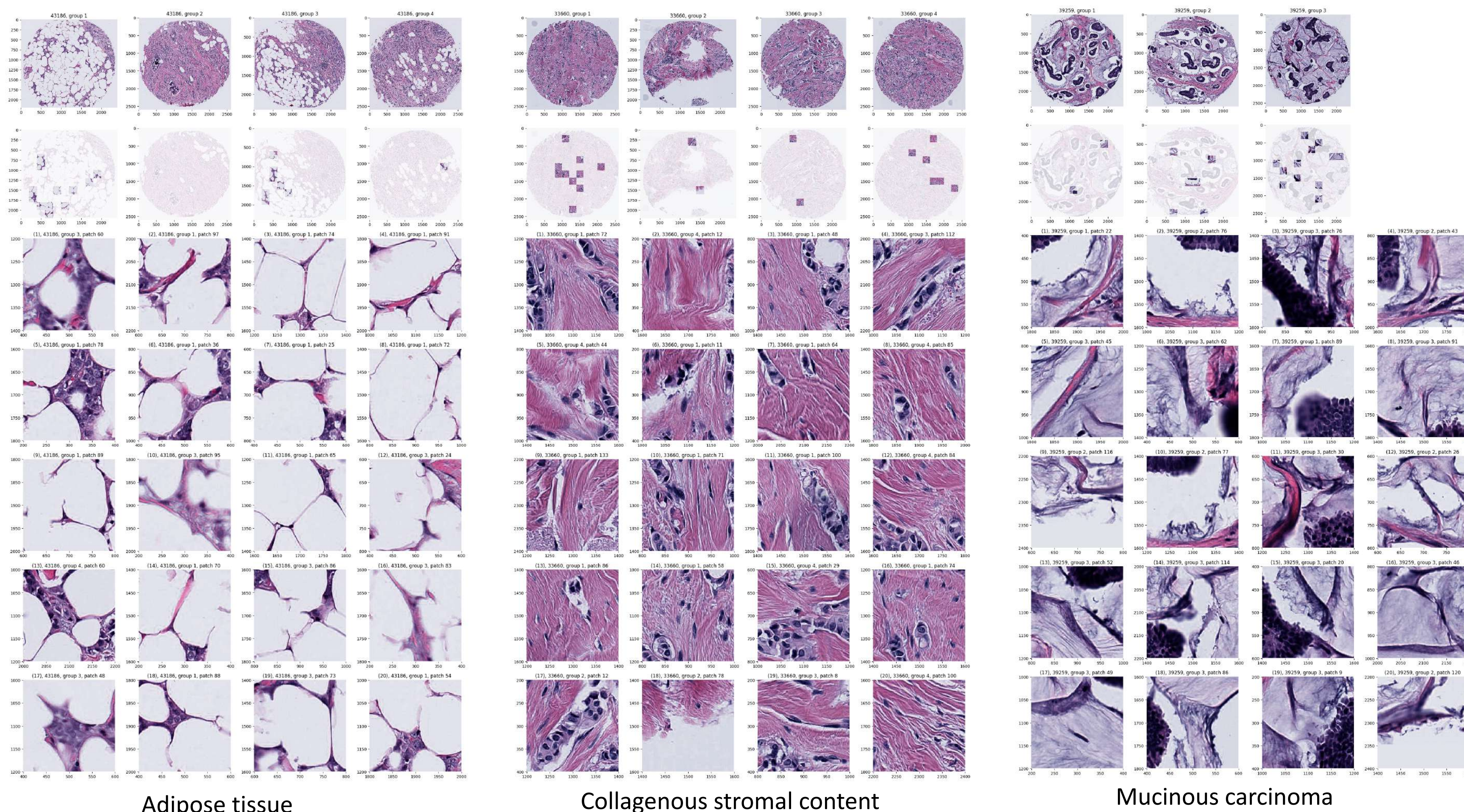- Representative patches are then selected by via projection onto individual loading vector

Adipose tissue

Collagenous stromal content

Mucinous carcinoma

Image individual components find signals which are known to be unrelated to PAM50
- E.g. fat content, stroma, mucinous carcinoma

Image and genetic joint

Showing
- PAM50 loadings vector for joint component 1
- Two representative subjects for joint component 1
- Saliency map visualization of single patch

Saliency map for first joint component

PAM50 loadings vector

Dense lymphocytic infiltrate
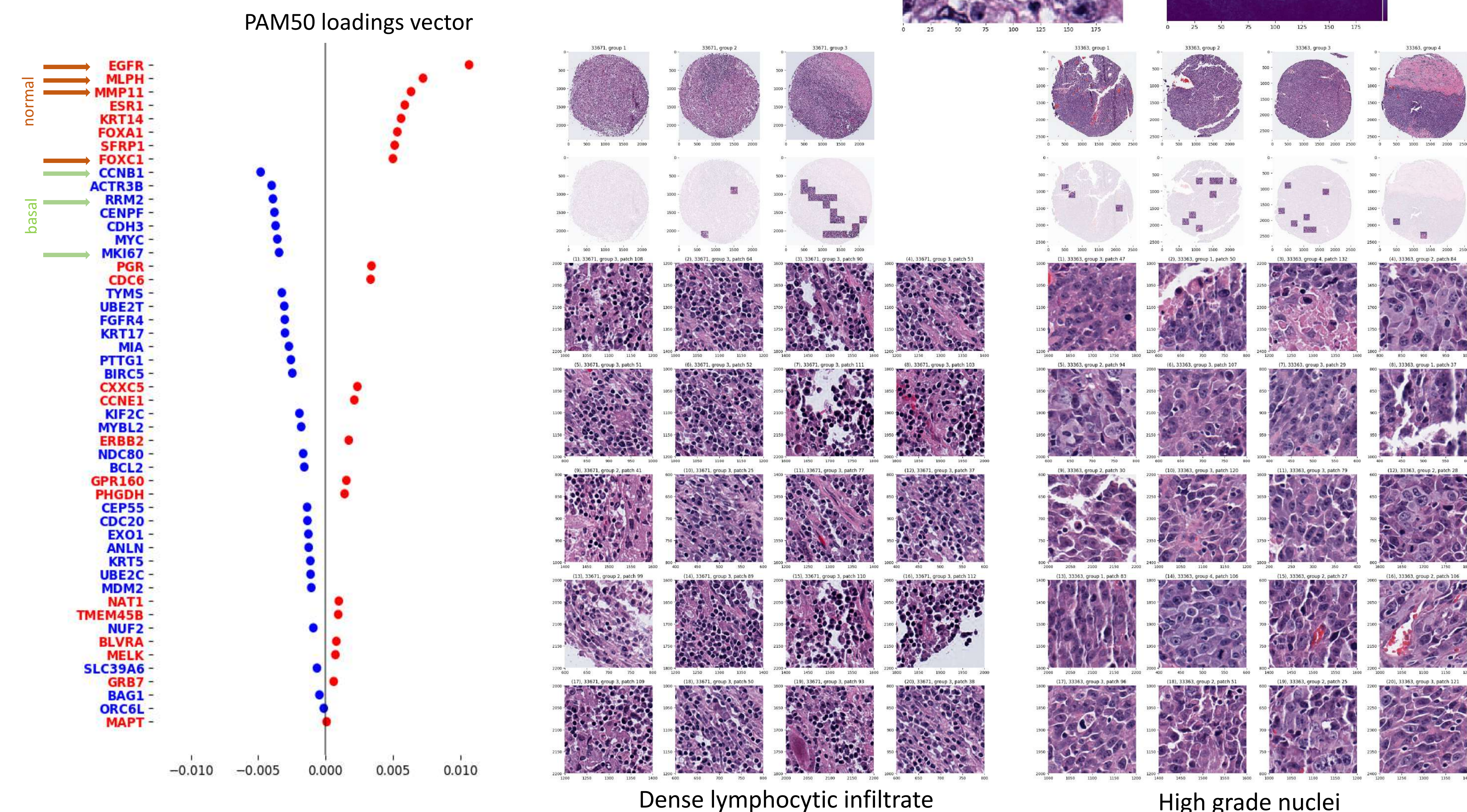
High grade nuclei

Joint components find signals which are known to be associated with both PAM50 genes and histopathology
- E.g. first joint component finds high vs. low grade (basal vs. normal genes, lymphocytic infiltrate, high grade nuclei)