Data

Each slide is associated with a number of different information files. The primary data file is a tab-delimited matrix, with a row for each detected cell on the slide, and columns representing the variables. These files are named “\*\_cell\_seg\_data.txt”, e.g. “halle validation\_\_1441-13\_HP\_IM3\_0\_[19680,14583]\_cell\_seg\_data.txt”.

I processed each file into a reconstructed cell-location matrix using the pixel-coordinates and phenotype of each cell. The following phenotype mapping is used:

0 = no cell

1 = tumor (pdl1-), blue

2 = pdl1 (pdl1+), red

3 = foxp3, green

4 = CD8, yellow

5 = CD4, magenta

6 = macrophage (pdl1+), cyan

7 = other (e.g. muscle cell), white

8 = macrophage (pdl1-), orange

These matrices are subsequently stored as numpy binary files. Because each cell is now only recorded as a single pixel, information about the size and shape of the cell is lost. Alternatives to this approach are: 1) working directly on the original images, which would work best on multi-channel TIFs, 2) approximating cell size by mapping the cell pixels to a circle of constant size, 3) approximating cell size by mapping to a circle of size set by the cell area variable in the data files.

Next, the matrices were spatially segmented into tumor and stromal regions using K-nearest-neighbors with K=25. By distinguishing tumor and stroma, this allows us to add regional features and calculate cell densities. In addition, many slides contain regions that were not processed by the Enform machine. The segmentations were filtered using simple image processing techniques to remove these regions and prevent density calculations from being affected by the extra area.

In addition, as an alternative boundary, I directly scraped the original tumor/stromal segmentation images returned by the Enform software, which are stored as “\*\_image\_with\_tissue\_seg.jpg”. This was done using color range matching. A visual comparison of the two methods follows:

In the following analyses, I attempt to predict pdl1 expression in tumor cells at differing spatial resolutions. The predictor covariates to be used come from the locations and phenotypes of the other cells. To include information from other tumor cells in the covariates (without giving away the pdl1 expression of the cells), I combined pdl1- and pdl1+ as a single tumor phenotype.

Whole slide pdl1 expression

Our first question of interest is whether the biopsy from a given patient contains tumor cells expressing pdl1. The prediction target is the percentage of all tumor cells with pdl1+. Using the phenotype designations of Table 1, this is computed as #pdl1 / (#pdl1 + #tumor). Our data is split into slides, with generally 3 slides per patient, but we wish to extrapolate from asking “Does the given slide contain pdl1” to “Does this patient’s tumor express pdl1?”. As shown in Fig \_\_, there is high concordance of pdl1 expression between different slides from the same patient, allowing us to study this question.

At the slide level, the cell-location and segmentation matrices were used to compute the following sets of features:

1. density of each phenotype within the tumor region,
2. density of each phenotype within the stromal region,
3. local pairwise cell-count correlations between phenotypes,
4. Features representing basic interactions between phenotypes, made by taking polynomial combinations of the features in 1) and 2) with degree 2.

The 3rd group of features were generated by splitting the slide into K x K tiles and taking the cell counts of each phenotype within each tile as observations. Pearson correlations were computed pairwise between phenotypes using these counts. This can be seen as a measure of distribution similarity between phenotypes at varying scales (perhaps mutual information/KL distance instead of correlation would serve as better predictors for future analysis).

I also very briefly experimented with computing slide-level texture metrics and co-occurrence between cells. I used the log of the features in 1), 2), and 4).

We have relatively few patients (around 120) so I tested the models using grouped 10-fold cross-validation. Grouping refers to splitting the data so that slides from the same patient are in the same fold. Since there is large similarity between slides from the same patient, having those slides in both a train and test set leads to over-optimism. Because we have a large number of features, I used ridge regression (L2-regularized linear regression) as the method on each fold.

Accuracy metrics for the method are in Table \_\_ and a plot of predictions are shown in Fig. \_\_\_.

Notes for improved analysis:

* Linear regression is a problem for the target variable which falls in the range [0,1]. In the future, the features should be passed in a generalized linear model with binomial/beta family. Observations are highly clustered around 0 and 1 so a beta distribution would be ideal.
* Alternatively, it may make more biological sense to model as a mixture distribution with a group with “low” pdl1 expression and another group with “high” pdl1 expression. I haven’t thought about this in detail but I think it could be modeled using a random effects model using the EM algorithm, and can be done with the flexMix package in R.
* Most importantly, I have only scratched the surface of the space of potentially useful features involving “interactions” between phenotypes (products/ratios via polynomial features and co-distributions using tiling). There are many other conceivable features that could be useful, from texture and co-occurrence, to modeling cells using random graphs. Ideally a logical and principled approach could be devised to incorporate these while also permitting biological interpretation. I also plan to read into graph convolutional methods to potentially generate complex features without requiring as much supervised engineering.

Localized pdl1 expression

The goal of this section is to attempt to predict pdl1 expression in tumor cells at a local spatial level. The target of the experiment is, given a disc of diameter 150px, the percentage of tumor cells within the disc that are pdl1+. The overall approach to generating the data is as follows:

Each slide is rectangular with dimension 1040x1392px. The slide was partitioned into 150x150px tiles. Within each tile I inscribe a circular mask with diameter 150px and count the cells within. I repeat the process to generate a sequence of concentric discs of increasing diameter: 212, 300, 424px. The diameters were chosen so that the area within each concentric region is constant. The predictor variables collected are the counts of each cell phenotype within each concentric region, i.e. within the 150px disc, outside the 150px disc but within the 212px disc, and so on.

The number of such observations taken from each slide is constrained by 3 factors:

1. To avoid bias, the 424px disc must fit within the boundaries of the slide.
2. Further, the 424px disc must not overlap significantly (here defined to be >10%) with areas not processed by the Enform machine, here assessed using the tumor/stromal segmentation files.
3. Finally, observations with under 10 tumor cells in the 150px disc were removed to ensure the observed area is within a tumor region.

Following these constraints resulted in 3228 observations from the 314 slides. I split these observations into 70%/15%/15% training/validation/test sets. I fit a set of models on the training set and observed their performance on the validation set. These models are:

1. Binomial generalized linear model (GLM): In terms of statistical analysis, this is an appropriate extension of linear regression because the outcome variable lies in [0,1]. The model fits a conditional binomial distribution, with the outcome variable representing the ratio of successes (# pdl1+ tumor cells) to total trials (# tumor cells). In terms of out-of-sample prediction, this model performed identically to simple linear regression and ridge regression.
2. Random forest model
3. 2-cluster binomial mixture model: I hypothesized the data could represent a mixture of two groups, one with low pdl1 and one with high pdl1 expression. This is because the distribution of pdl1 expression is highly bimodal and clustered at 0 and 1. Using the EM algorithm and the binomial GLM likelihood, the observations are split into a binomial mixture model. If we knew the latent cluster identifications *a priori*, the model does very well on the validation set, with MSE=0.07 and correlation=0.90. However, the clusters can only be identified using both the covariates and the outcome variable, so this is not possible out-of-sample. Thus I tried to infer the out-of-sample cluster identities using a classification algorithm on the covariates (here I tested logistic regression and random forest, and random forest did much better).

The result of 3) suggests that predicting pdl1 expression can be reduced to a classification problem (i.e. low vs. high pdl1 expression), because once the classification is known, the subsequent regression is accurate. Unfortunately in either case, the collected covariate data doesn’t produce great classification or regression, although we do see evidence of significant associations.

Finally, I think the existing feature space (cell counts in each region) alone is too limited and loses too much information. May turn out to be much better to treat it as more of a computer vision problem, passing all the pixels directly into convolutional layers and using a deep learning framework rather than just using cell counts.