DeepLIFT Feature Exploration of a Cancer Cell-of-Origin Classifier

BCB430Y1 Final Lab Presentation - Mar. 27, 2019

Yoonsik Park, Gurnit Atwal, Quaid Morris

Background

Methods

Exploration

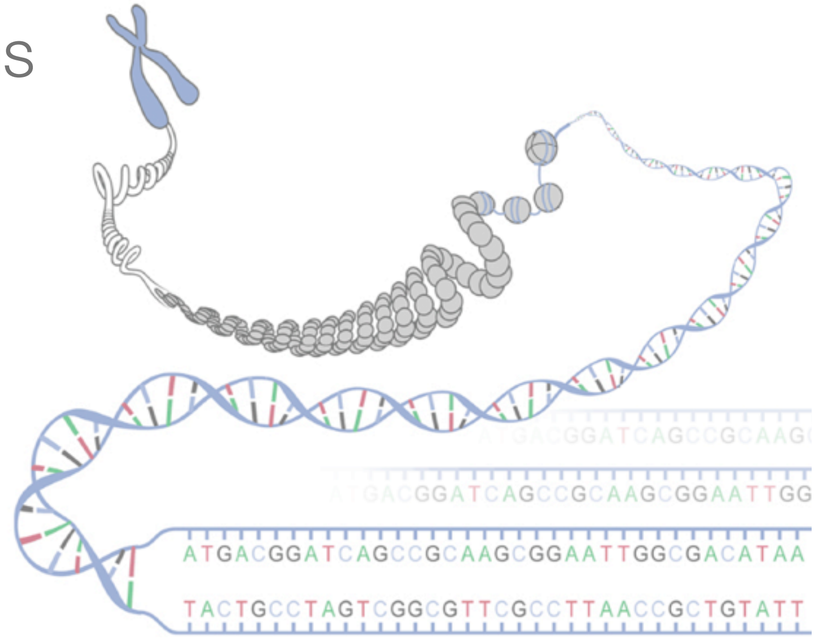
Background

Methods

Exploration

Cancer is caused by deregulation of genes

- Mutations and epigenetic changes to genes result in the disruption of normal cell death, division and proliferation
- Deregulation of oncogenes + tumour suppressor genes
- Cancers have driver mutations + passenger mutations



Mutation distribution is predictor for cancer cell type of origin

- The spatial distribution of mutations has been shown to be an excellent predictor of cell type of origin for different cancer types (Atwal, 2019)
  - Example: mutations in chr2, 17,000,000-18,000,000
- Furthermore, the distribution of base substitution classes has provided mutational signatures of cancer (Alexandrov, 2013)
  - Example: mutations of TCC --> TGC
- However, little insight into what features are allowing the classifier to achieve its performance

DeepLIFT provides insight into important features

- Stands for **Deep** **L**earning **I**mportant **F**ea**T**ures
- Python package created with a focus on interpretable deep learning
- Gives importance scores for features used to classify an example, for every class
- Back-propagates importance values by comparing against a reference example

Background

Methods

Exploration

Generate DeepLIFT Scores

- DeepLIFT scores were generated for one fold of the cancer classifier
- Feature importances were generated for every test example, against every class
- Importance values were then averaged across a class
- DeepLIFT scores were generated for each of the 24 cancer classes

Select Top Features

- DeepLIFT Feature Score Interpretation:
  - Negative -> contributes to lowering class output
  - Positive -> contributes to increasing class output
  - Zero -> no contribution for class
- Therefore, three sets of top 30 features were chosen by:
  - maximum of DeepLIFT scores
  - minimum of DeepLIFT scores
  - maximum of the absolute value of DeepLIFT scores

Background

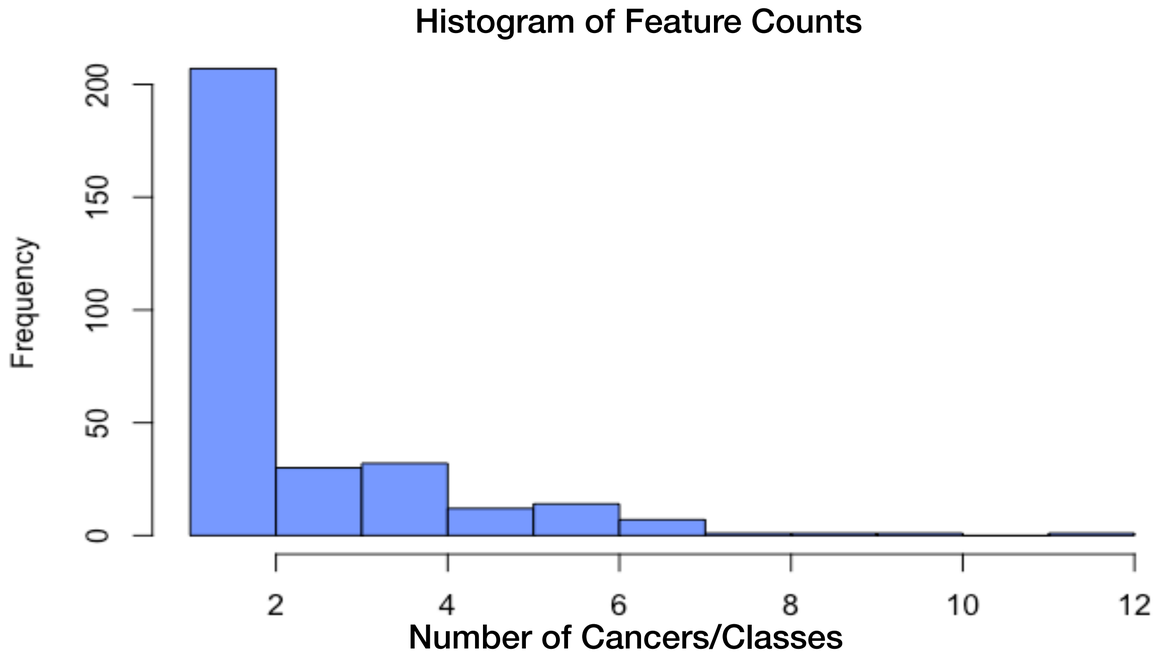
Methods

Exploration

Initial Findings

- For all 24 cancer classes, only spatial mutation density features (not base-substitution classes) were in the top 30 features for both negative and positive DeepLIFT scores

- Most features are unique to one class



Features most common across cancer types

Feature	Count
chr14.106	**12**
chr8.111	**10**
chr2.140	**9**
chr14.84	8
chr4.62	7
chr5.19	7
chr5.24	7
chr7.52	7
chr11.40	7
chr11.132	7
chr13.64	7
chr1.190	6
chr2.80	6

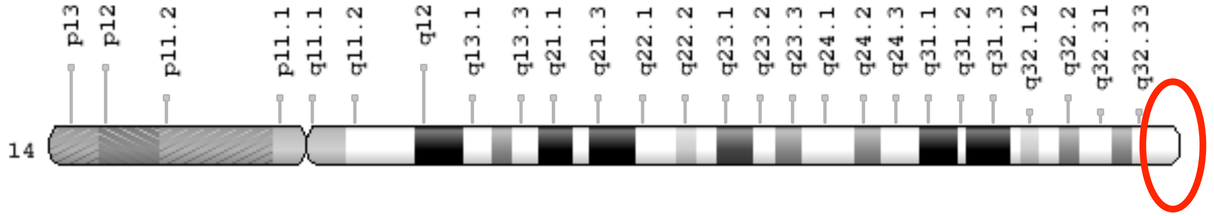
...

Feature	Count
chr5.2	6
chr5.3	6
chr5.26	6
chr7.41	6
chr7.62	6
chr7.68	6
chr8.5	6
chr9.10	6
chr12.18	6
chr13.93	6
chr18.65	6
chr20.41	6

Feature 1: `chr14.106`

POSITIVE	NEGATIVE	Less Contribution
Lymph-BNHL	**CNS-Medullo**	Bone-Osteosarc
Lymph-CLL	**CNS-PiloAstro**	Breast-AdenoCA
Myeloid-MPN	**Eso-AdenoCA**	CNS-GBM
Thy-AdenoCA	**Kidney-ChRCC**	ColoRect-AdenoCA
	Kidney-RCC	Head-SCC
	Lung-AdenoCA	Liver-HCC
	Lung-SCC	Ovary-AdenoCA
	Panc-AdenoCA	Panc-Endocrine
	Uterus-AdenoCA	Prost-AdenoCA
		Skin-Melanoma
		Stomach-AdenoCA

Feature 1: chr14.106

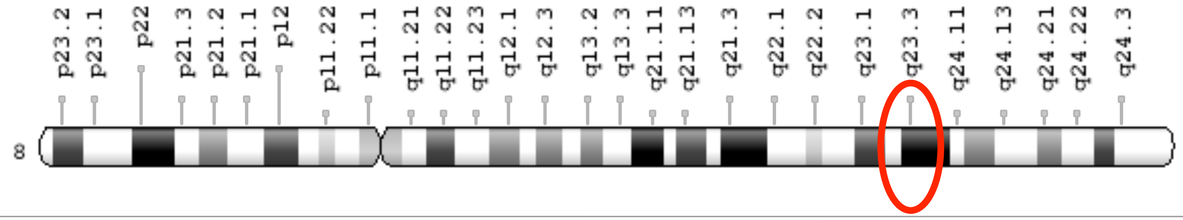


- Segment contains: ~123 IGH-V genes, 5 lncRNAs, 1 miRNA, 3 pseudogenes
- Subset of the immunoglobulin heavy locus in humans, for VDJ recombination and somatic hypermutation
- Such an important predictor for chronic lymphocytic leukemia (CLL) survival and treatment (Crombie, 2017), that it has its own name: IGHV%, for percent deviation
- miR-5195-3p (miRNA) also inhibits proliferation of human bladder cancer cells (Jiang, 2017)

Feature 2: `chr8.111`

POSITIVE	NEGATIVE	Less Contribution
ColoRect-AdenoCA	**CNS-Medullo**	Bone-Osteosarc
Head-SCC	**CNS-PiloAstro**	Breast-AdenoCA
Liver-HCC	**Kidney-ChRCC**	CNS-GBM
Stomach-AdenoCA	**Myeloid-MPN**	Eso-AdenoCA
	Ovary-AdenoCA	Kidney-RCC
	Thy-AdenoCA	Lung-AdenoCA
		Lung-SCC
		Lymph-BNHL
		Lymph-CLL
		Panc-AdenoCA
		Panc-Endocrine
		Prost-AdenoCA
		Skin-Melanoma
		Uterus-AdenoCA

Feature 2: `chr8.111`

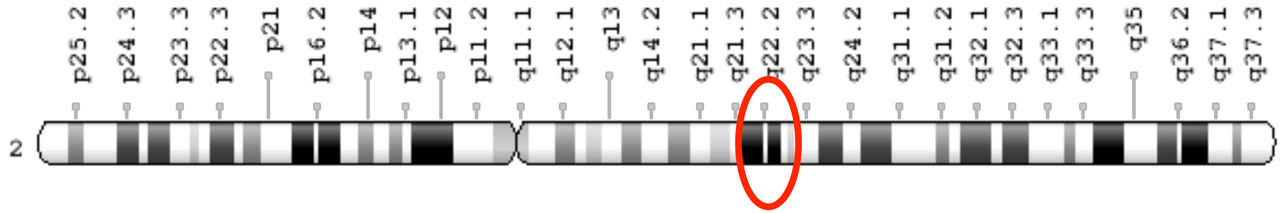


- Segment contains: 2 pseudogenes, 9 lncRNAs
- Long non-coding RNAs (lncRNAs) are the majority of genomic products
- Multiple roles of long non-coding RNA in cancer (Sanchez Calle, 2018) e.g.
  - Regulation of tumor suppressors
  - Regulation of tumor drivers
  - Chromatin remodeling

Feature 3: `chr2.140`

POSITIVE	NEGATIVE	Less Contribution
Eso-AdenoCA	**Bone-Osteosarc**	CNS-Medullo
Liver-HCC	**Breast-AdenoCA**	CNS-PiloAstro
Lung-SCC	**CNS-GBM**	ColoRect-AdenoCA
Panc-AdenoCA	**Lymph-BNHL**	Head-SCC
Stomach-AdenoCA	**Myeloid-MPN**	Kidney-ChRCC
	Skin-Melanoma	Kidney-RCC
	Thy-AdenoCA	Liver-HCC
		Lung-AdenoCA
		Lymph-CLL
		Ovary-AdenoCA
		Panc-Endocrine
		Prost-AdenoCA
		Uterus-AdenoCA

Feature 3: `chr2.140`

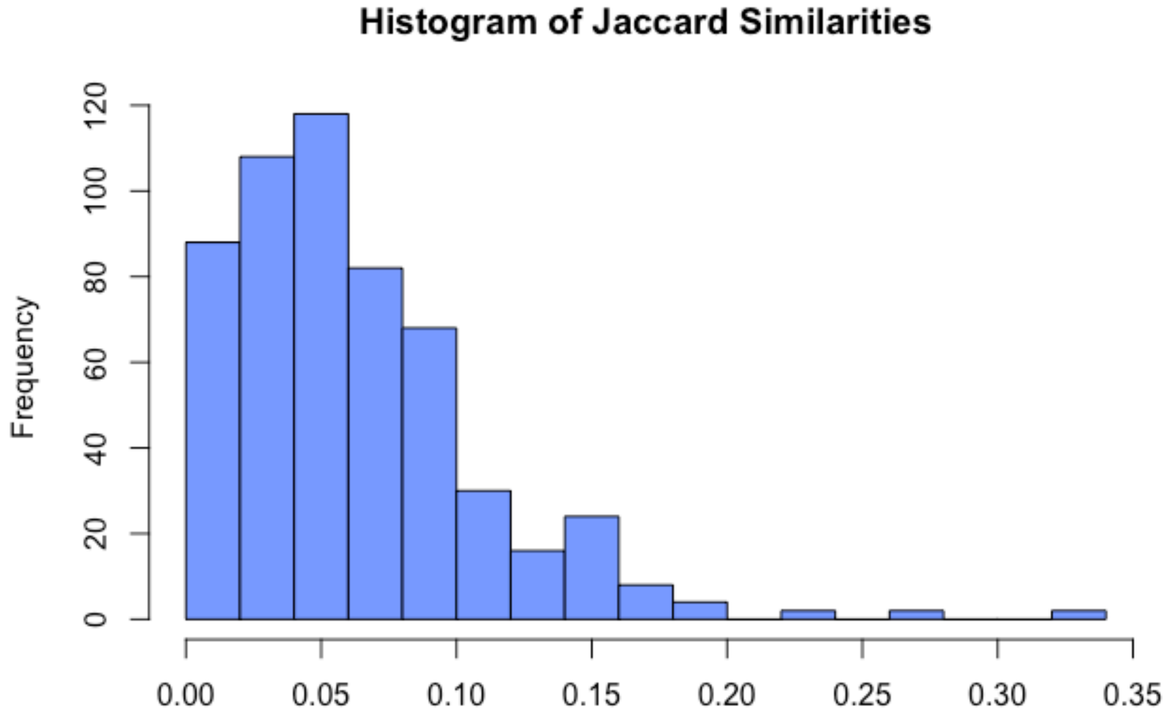


- Segment contains: LRP1B, 1 lncRNA, 1 miRNA, 3 pseudogenes
- LRP1B: low-density lipoprotein (LDL) receptor-related protein
  - Potentially on cell surface and interacts with ligands
  - Mutations have high prevalence in lung, renal, thyroid cancers
  - LRP1B was reported amongst the top 10 most significantly deleted genes across 3312 human cancer specimens (Beroukhim, 2010)

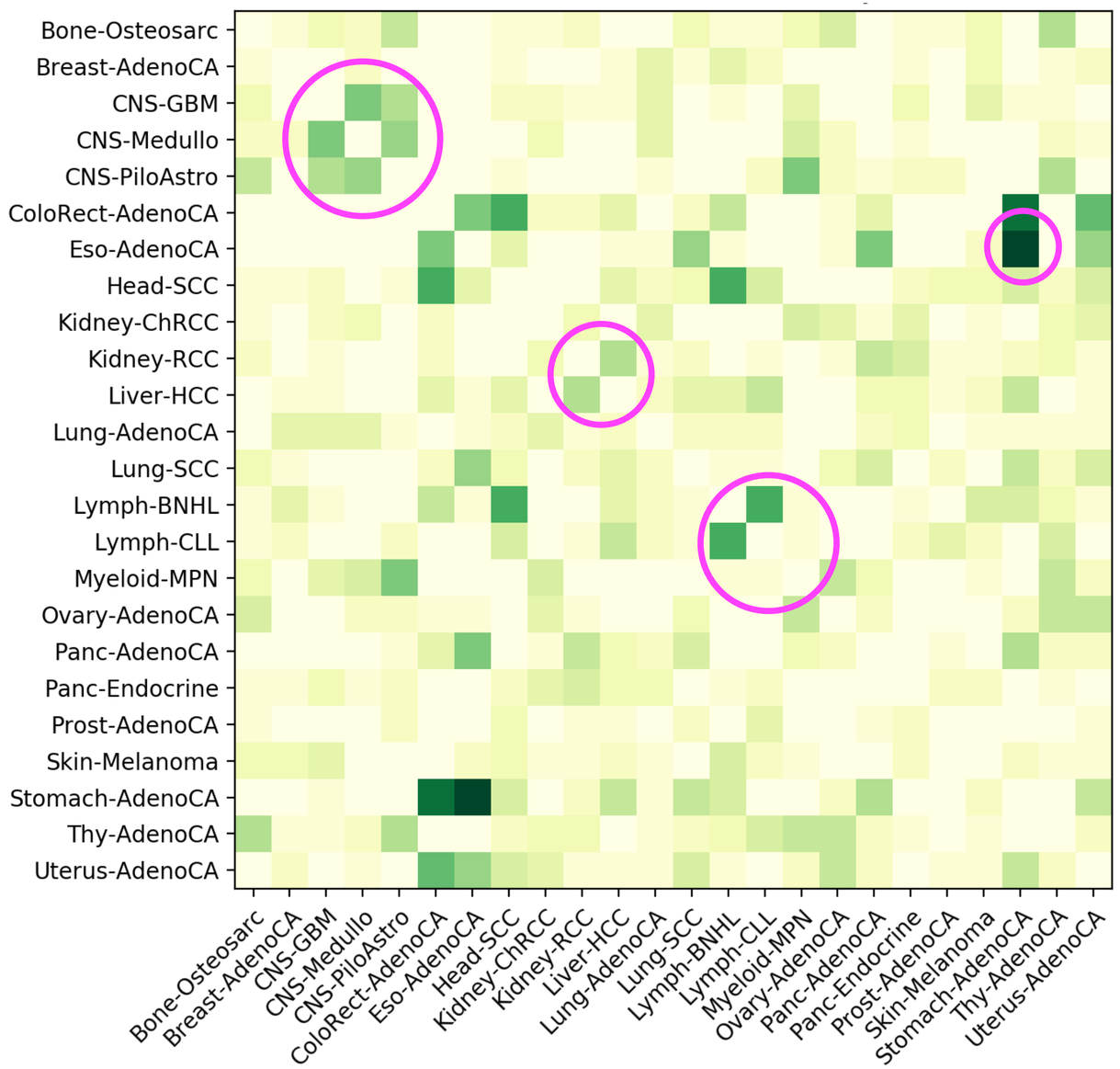
Jaccard Similarity of Features Between Cancer Classes

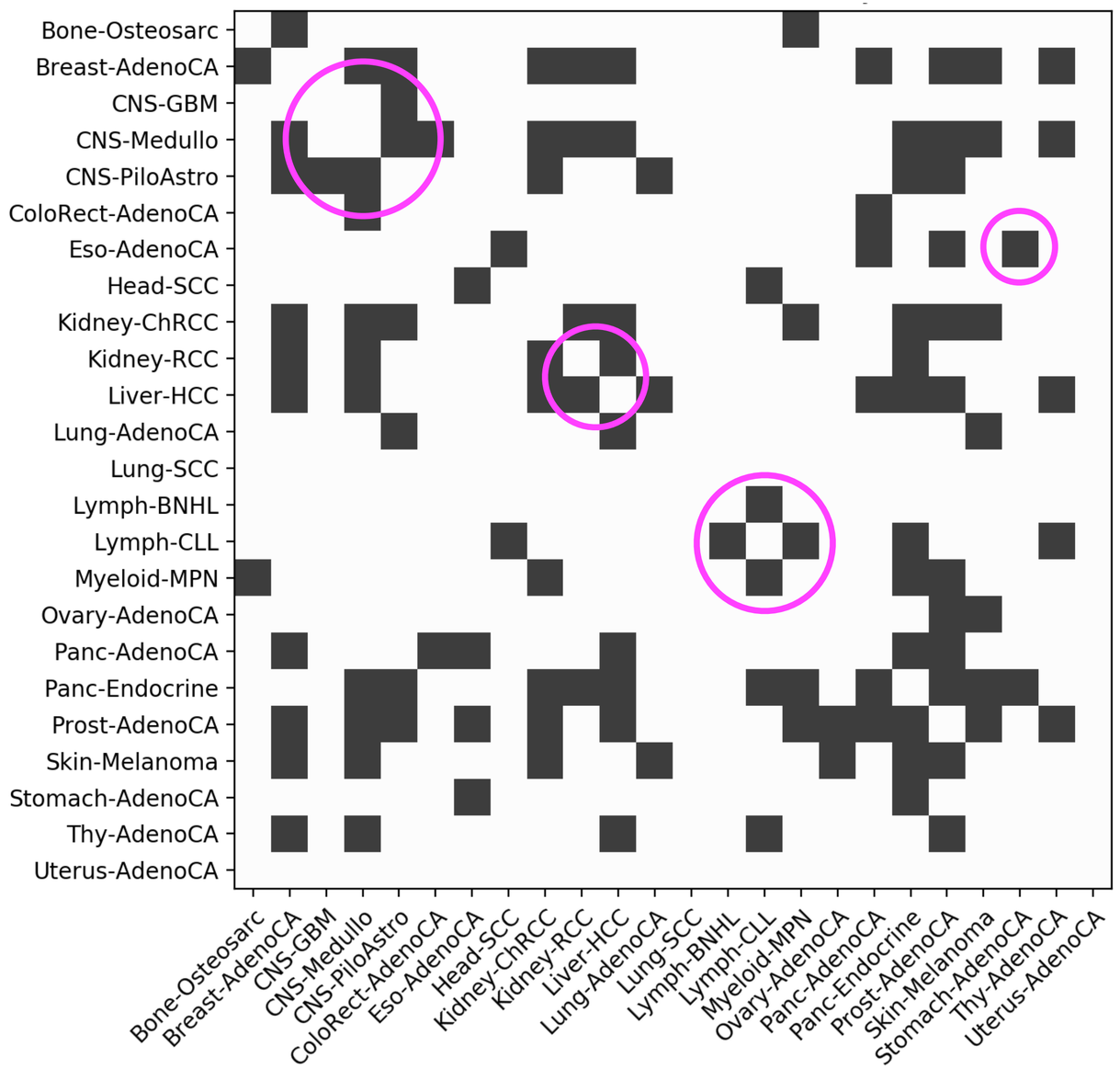
- Each Cancer Class has a Set of Features
- Compare Pairwise Similarity using Jaccard Index

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

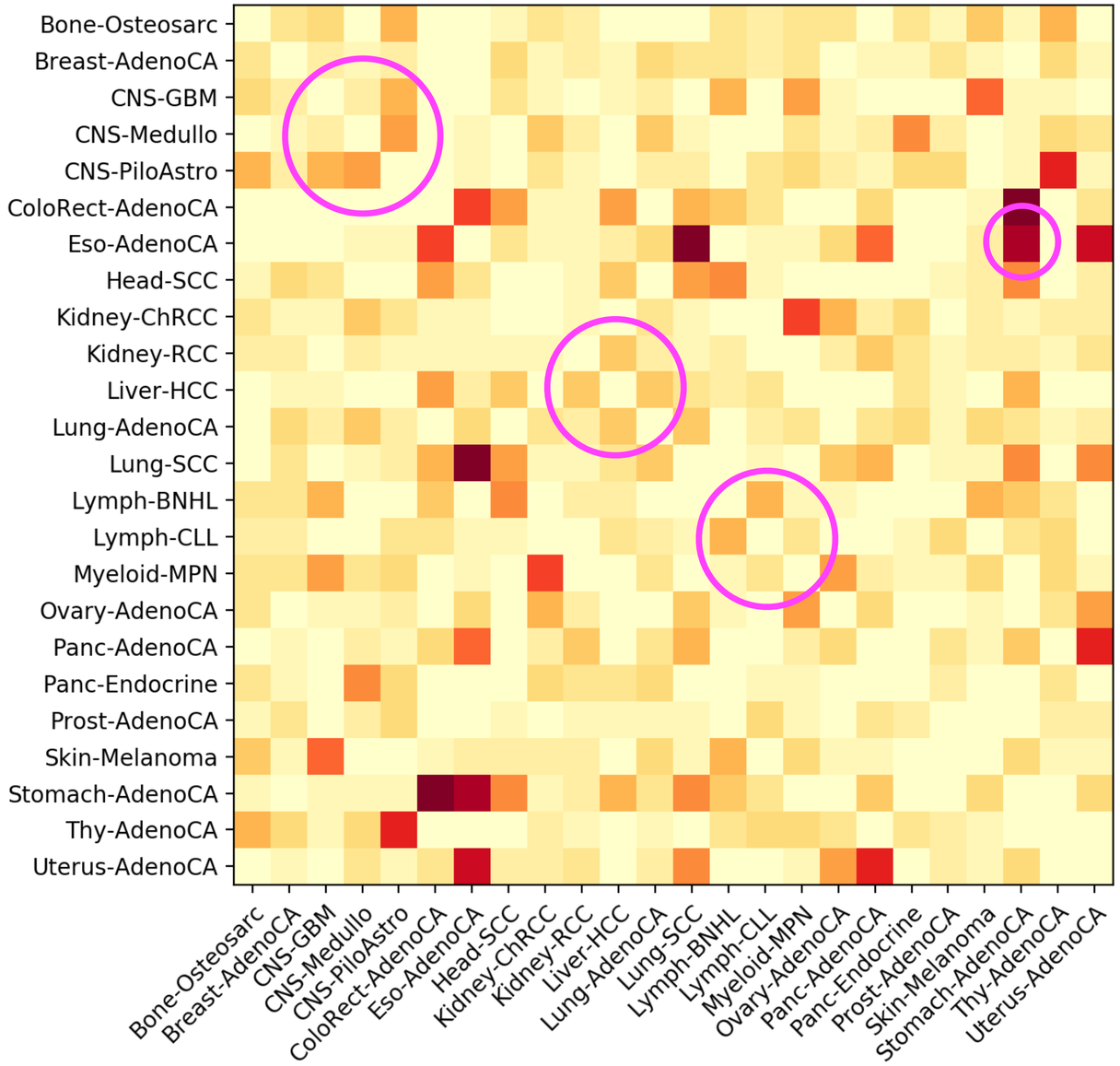


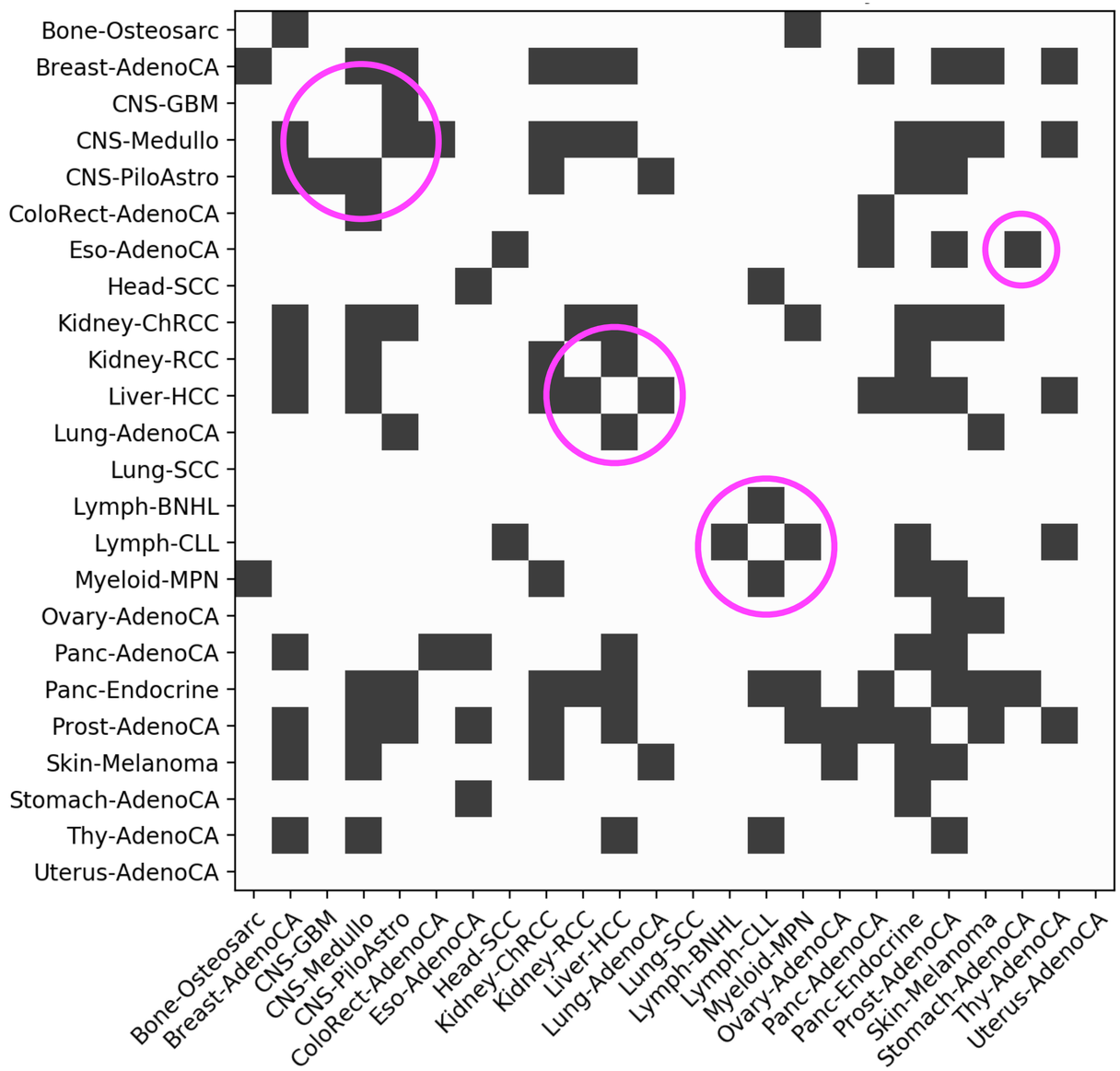
Positive Features Similarity Matrix vs Modified Confusion Matrix





Negative Features Similarity Matrix vs Modified Confusion Matrix





Concluding Remarks

- DeepLIFT has found multiple interesting regions that can be investigated further for biological significance
- Some cancer types that the classifier finds difficult are seen to have a higher Jaccard index for features
- In the future, may want to look at segments in the context of chromatin state and see if the mutations are from epigenetic silencing or driver mutations

Acknowledgments

- Gurnit Atwal
- Quaid Morris
- Morris Lab

Segment	FeatureID	Count
chr14.106	2315	12
chr8.111	1509	10
chr2.140	390	9
chr14.84	2293	8
chr4.62	755	7
chr5.19	904	7
chr5.24	909	7
chr7.52	1290	7
chr11.40	1863	7
chr11.132	1955	7
chr13.64	2157	7