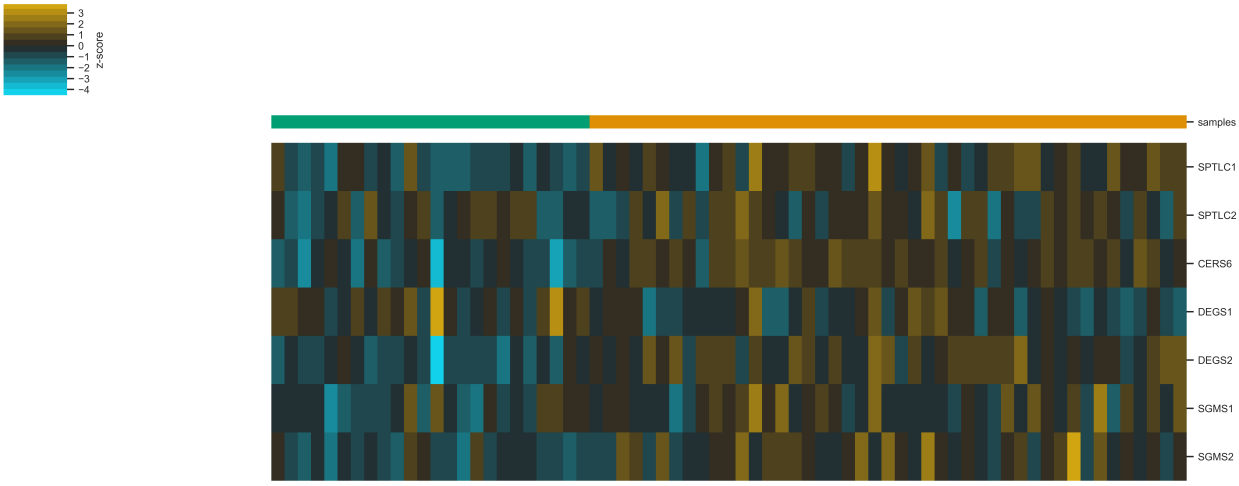
Analysis scripts for Li, et. al.

Figure Legends

Microarray expression values were z-score scaled. Samples were clustered via calculating the Euclidean distance between centroids. See (1) for raw data curation and (2) for the associated analysis code for this manuscript. The data was originally generated as part of (3) and (4) and can be accessed using the GEO identifiers GSE20916 and GSE8671 (5).

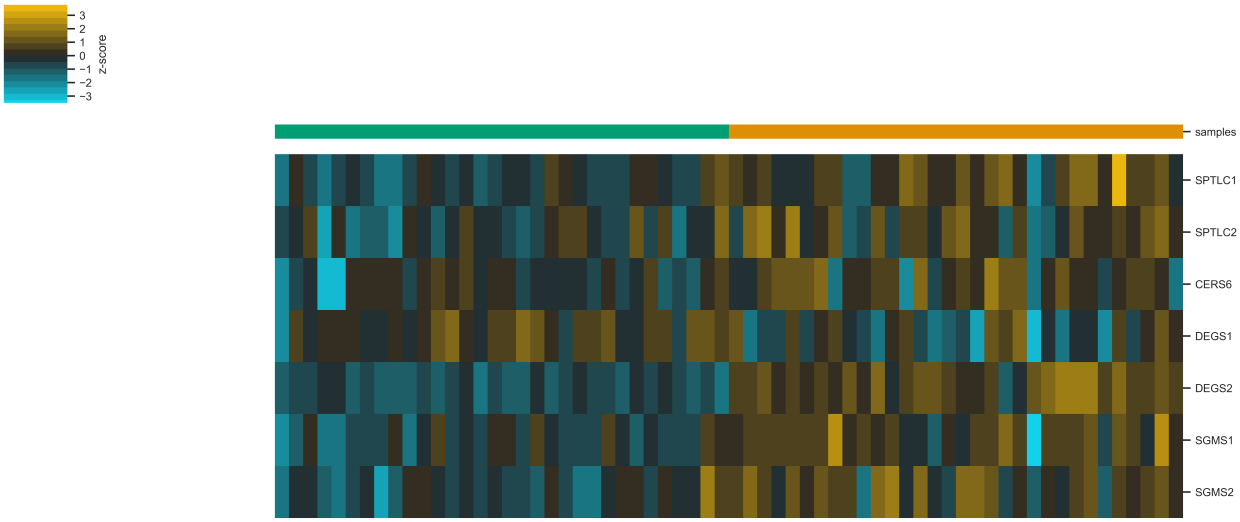
GSE20916



Stats

Gene	*d*	BH FDR
CERS6:	1.62 (Very large)	3.370457762577637e-07
DEGS1:	-0.6 (Medium)	0.02372965901312005
DEGS2:	1.59 (Very large)	2.0910343687627823e-07
SGMS1:	0.68 (Medium)	0.01840974024266771
SGMS2:	1.06 (Large)	0.00040751185719631387
SPTLC1:	1.37 (Very large)	1.2464218423931629e-05
SPTLC2:	0.24 (Small)	0.4488039915916772

GSE8671



Stats

Gene	*d*	BH FDR
CERS6:	0.65 (Medium)	0.010036555418121696
DEGS1:	-0.76 (Medium)	0.014659715354925305
DEGS2:	2.59 (Huge)	2.2571193260020697e-09
SGMS1:	1.04 (Large)	1.588817159913001e-05
SGMS2:	1.12 (Large)	0.0001774572757259638
SPTLC1:	1.09 (Large)	5.656634138044894e-05
SPTLC2:	0.75 (Medium)	0.012654179495407347

Methods

Microarray expression data were accessed and analyzed as described in (1). Briefly, human microarray data were accessed from the GEO database GSE20916 and GSE8671 (5) under the inclusion parameters of "normal" tissue and non-cancer "adenoma" tissue. Multimapping probes were dropped and probe sets mapping to the same gene were collapsed and averaged. Heatmaps were generated using XPRESSplot (v0.2.5) (6). GEO datasets were parsed using the GEOparse package (v2.0.2) (https://pypi.org/project/GEOparse/). Gene set normalization was performed using scikit-learn (v0.23.2) (7). Benjamini-Hochberg (8) corrected p-values were calculated using SciPy (v1.6.0) (9) and statsmodels (v0.12.1) (10). Cohen's *d* effect sizes were calculated using the following equation:

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s} = \frac{\mu_1 - \mu_2}{s}.$$

Effect sizes were scaled as follows (11-12):

Effect size	*d*	Reference
Very small	0.01	[12]
Small	0.20	[11]
Medium	0.50	[11]

Effect size	*d*	Reference
Large	0.80	[11]
Very large	1.20	[12]
Huge	2.0	[11]

Statistics and figures related to processing the heatmaps and GEO database datasets were performed in Python (v3.8.6). Data processing and analyses can be interactively replicated using Jupyter Notebook (https://jupyter.org) at (13).

References

[1] Bensard CL, Wisidagama DR, Olson KA, Berg JA, Krah NM, Schell JC, Nowinski SM, Fogarty S, Bott AJ, Wei P, Dove KK, Tanner JM, Panic V, Cluntun A, Lettlova S, Earl CS, Namnath DF, Vázquez-Arreguín K, Villanueva CJ, Tantin D, Murtaugh LC, Evason KJ, Ducker GS, Thummel CS, Rutter J. Regulation of Tumor Initiation by the Mitochondrial Pyruvate Carrier. Cell Metab. 2020 Feb 4;31(2):284-300.e7. doi: 10.1016/j.cmet.2019.11.002.
[2] https://github.com/j-berg/li_2021
[3] https://doi.org/10.1371/journal.pone.0013091
[4] https://doi.org/10.1158/1541-7786.mcr-07-0267
[5] https://www.ncbi.nlm.nih.gov/geo/
[6] Berg JA, Belyeu JR, Morgan JT, Ouyang Y, Bott AJ, Quinlan AR, Gertz J, Rutter J. XPRESSyourself: Enhancing, standardizing, and automating ribosome profiling computational analyses yields improved insight into data. PLoS Comput Biol. 2020 Jan 31;16(1):e1007625. doi: 10.1371/journal.pcbi.1007625.
[7] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E. Scikit-learn: Machine Learning in Python (2011). Journal of Machine Learning Research, vol 12.
[8] Yoav Benjamini and Daniel Yekutieli. The Control of the False Discovery Rate in Multiple Testing under Dependency (2001). The Annals of Statistics, Vol. 29, No. 4. https://www.jstor.org/stable/2674075
[9] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, CJ Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E.A. Quintero, Charles R Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. (2020) SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods, 17(3), 261-272.
[10] Seabold, Skipper, and Josef Perktold. “statsmodels: Econometric and statistical modeling with python.” Proceedings of the 9th Python in Science Conference. 2010.
[11] Cohen, Jacob (1988). Statistical Power Analysis for the Behavioral Sciences. Routledge. ISBN 978-1-134-74270-7.
[12] Sawilowsky, S (2009). "New effect size rules of thumb". Journal of Modern Applied Statistical Methods. 8 (2): 467–474. doi:10.22237/jmasm/1257035100.
[13] https://doi.org/10.5281/zenodo.4567545

To reproduce the analyses from these scripts:

The following example will show how to install and run the analyses on a *nix OS.

1. Download Conda and install.
2. Download this repository.
3. Unzip the folder and navigate to the folder using the command line. For example, if you downloaded the zip file to your Desktop:

```
cd ~/Desktop/
unzip li_2021-main.zip
cd li_2021-main
```

4. Create a conda environment:

```
conda env create -n li_analysis -f requirements.txt
conda activate li_analysis
conda activate jupyter
pip install GEOparse
```

5. Launch Juypter Notebook (from within the `li_2021-main` directory):

```
jupyter notebook
```

Requirements:

- `Python3`
- `Pandas`
- `NumPy`
- `Matplotlib`
- `Seaborn`
- `Scikit-Learn`
- `XPRESSplot`
- `GEOparse`

Add to citations

- `scipy`
- `statsmodels`