Included in this repository is all of the code and raw data necessary to replicate the analyses of "A pan-cancer analysis of prognostic genes", with the exception of the mrna expression files which are too large to store on GitHub. Those files can be freely downloaded from https://tcga-data.nci.nih.gov/tcga/. You will be expected to put these files in tcga\_data/cancer/mrna/. The exact clinical data used is already present in tcga\_data/cancer/clinical. If you replace these clinical files with more recent files from TCGA there is no guarantee the files will get parsed correctly. Replacing the "FILE\_SAMPLE\_MAP.txt" file present at /tcga\_data/cancer with a more up to date one should not break any code.

To replicate the analysis start in the cox\_regression folder. Each cancer has its own 'cox\_regression.py' script. This script was run with python 2.7 and requires rpy2 and NumPy to be installed. If you downloaded the mrna files and placed them in the correct folder this code should run from the command line with "python cox\_regression.py". Depending on the cancer this script could take ~30 min to run and several GB of RAM.

The script will generate a list of cox coefficients and pvalues for every gene that met the expression cutoff, and save the results to "coeffs\_pvalues.txt", which is already present in the folder. By default, the code will not save the "final\_genes.txt" file, which saves the expression values of all the genes in a convenient format for python. This file would be needed if you are interested in reproducing Fig. 1b. You can uncomment out the code if you want this file saved. The "final\_genes.txt" file is not present in the repository due to file size limits.

The "adjusted\_pvalues.py" script will BH adjust the p-values in "coeffs\_pvalues.txt", and write a file "coeffs\_pvalues\_adjusted.txt".

The "normalizing\_coeffs.py" script will read "coeffs\_pvalues\_adjusted.txt" and sigmoidal normalize the coefficients and write a file "coeffs\_normalized\_pvalues\_adjusted.txt"

The "patient\_info.py" script will get some summary information about the patients included in the study and write a file "patient\_info.txt".

The "msigdb.py" file will print out the 250 most significant protective genes, and the 250 most significant harmful genes. These are the genes which were used for msigdb gene set analysis. The resulting output files from msigdb are saved as good\_overlap and bad\_overlap.

The code for generating the four figures and one supplemental figure are present in the "figures" folder. Also included is the final figure pdf along with the inkscape svg file. The code for generating all of the tables is included in the "tables" folder.

During commenting of the code in preparation of submission to GitHub it is possible typos were introduced. Please report any errors or problems to omnesresnetwork@gmail.com.