The R script written merges clinical data with the already combined gene expressions, mutations, copy number variations, and methylation automatically. Automating the merging process allows for easier merging in the case of future renditions made to any of the data. The script will allow anyone with access to similar or the same data to combine their datasets too.

The datasets were merged based on the patient’s bcr (breakpoint cluster region protein) barcode. Editions had to be made to the clinical dataset for the barcodes to match exactly, due to a difference between hyphens and periods, the inclusion of the word “data”, and capitalization. Another edition made to clinical was to add .01 at the end of the barcode, because each barcode at least contained .01 (symbol for cancerous sample) at the end in the other dataset’s barcode. The other endings, .10 and .13, stood for normal samples and metastasized samples respectfully. This ensured that all the proper barcodes were given the additional variables from clinical, without any repeated values. By merging these two data sets, it allows for easier comparisons of the all the variables against each other.

An outer join was used to ensure no variables were lost while still merging identical barcodes together. The outer join was used through R’s merge function and setting the “all” parameter to true. By merging the clinical data with the others mentioned above, we introduced an additional 3719 variables.

