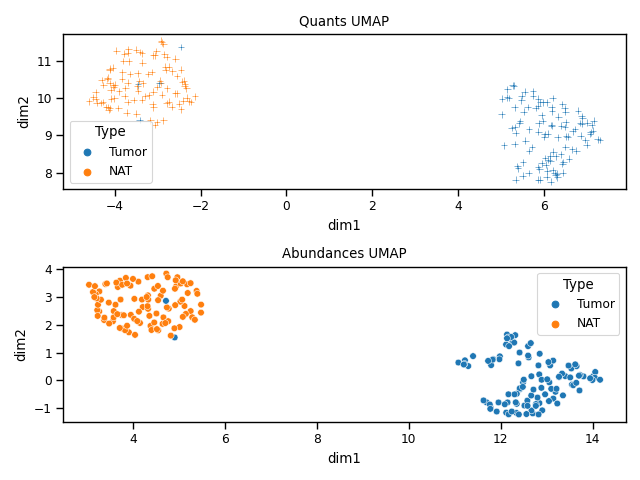
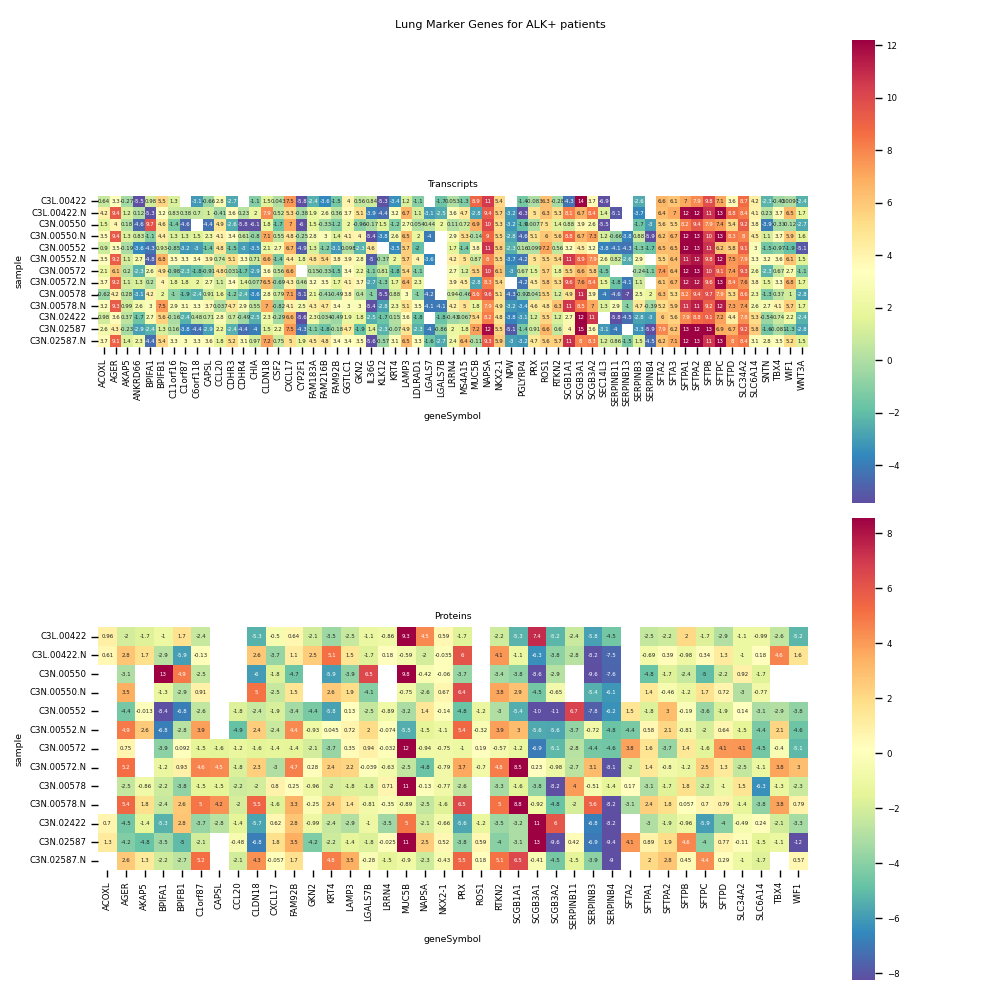
The general strategy I like to begin with when faced with a morass of data is to (after data wrangling) begin with sketching a global picture of the samples and their relationships with each other through standard clustering algorithms and dimension reduction techniques. Unsurprisingly, the largest source of variation is a sample’s status as either a tumor or NAT when looking at the first two principal components. The story appears similar when doing a 2D projection with UMAP, although it is curious two tumor samples in both modalities cluster with NATs.

Although using the entire transcriptomic and proteomic tell similar (and uninteresting) stories when looking at all of the samples, I hypothesized that a more interesting story could emerge from closely looking at a more intentionally chosen samples and genes. I was curious about the non-smokers who’ve had the misfortune of developing lung cancer. Sorting and filtering the metadata showed ALK-Fusions overrepresented in this cohort—all young (< 70) females with the same mutational signature ([COSMIC 5: unknown aetiology](https://cog.sanger.ac.uk/cosmic-signatures-production/images/v2_signature_profile_5.original.png)). Anaplastic lymphoma kinase (ALK) [can fuse with other genes to drive several cancers](https://molecular-cancer.biomedcentral.com/articles/10.1186/s12943-018-0776-2); ALK-positive cancers [represent 4% of lung cancers and the patient profiles are oddly young non-smokers](https://www.lung.org/lung-health-diseases/lung-disease-lookup/lung-cancer/symptoms-diagnosis/biomarker-testing/alk-lung-cancer).

After sub-selecting the samples, to get my bearings, [I pulled lung-specific genes](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4932457/pdf/FEB4-6-774.pdf) (analogizing them to “house-keeping genes”) and plotted both modalities for the ALK+ patients. In retrospect, perhaps the idea of looking at “standard gene expression” in cells nearby to tumors was a naive approach, because the data doesn’t quite look like I had expected it to (i.e., strong signal for most of the genes in all of the NATs).

That said, several intriguing profiles appeared. First, the SFTP\w\d genes show strong transcriptional signal in all of the ALK+ samples whereas the corresponding protein abundances show greater variance. Discordance between patterns of signal appear for AGER (loss of RAGE —> impaired adhesion), MUC5B (strong proteomic signal in tumor samples compared to NAT for all patients), and NAPSA (strong transcriptomic signal across all samples, [described as a marker for lung adenocarcinomas](https://www.genecards.org/cgi-bin/carddisp.pl?gene=NAPSA), protein abundances seem less conclusive). The proteomic data does not have a 1:1 correspondence with the transcriptomic data and points to more complicated mechanisms of control existing between the transcript, protein, and ultimate phenotype.

With additional time, several additional sensible (yet simple) approaches could be done: what genes show the largest difference in their transcriptomic vs. proteomic profile and are they enriched for any pathways or processes in particular; what do these profiles look like in healthy patients (to get better heuristics for judging aberrations); and of course looking at different putative driver mutation groups. The code is currently set up to generate investigatory plots for a given list of samples and genes, allowing for rapid (but targeted) exploration. Beyond me, but an idea that sounds appealing, would be something similar to approaches taken for [multi-model single cell data integration](https://pubmed.ncbi.nlm.nih.gov/34062119/) and “reducing” this dataset for a given population to its “anchors” (genes where the neighbors in expression and abundance space are the same), although as I type this I am wondering if I am just talking about Spearman correlation. Stratifying the ALK+ patients based on *which* gene they fuse to could also be interesting.

[ALK+ cancers exist in other tissues](https://www.uptodate.com/contents/anaplastic-lymphoma-kinase-alk-fusion-oncogene-positive-non-small-cell-lung-cancer) as well—it would be interesting to compare ALK+ lung adenocarcinomas to another tissue. Although treatments exist for ALK+ cancers, [resistance is an issue](https://www.frontiersin.org/articles/10.3389/fonc.2021.713530/full). Followup studies looking at the same patient and if resistance can be identified through transcriptomic and/or proteomic signal would be interesting