**TF identification**

to look for gain-of-binding site events, we will perform a SNVs analysis in regulatory elements proceeding as in [35]. In summary: we will query the COSMIC Cancer Gene Census for TFs (termed CGC-TFs). For these factors, we will search the JASPAR and TRANSFAC motif databases for motifs, construct a position-specific scoring matrix (PSSM), and determine the threshold PSSM value false-positive rate. Using the COSMIC SNV data for our cell line, we will filter for SNVs in regulatory regions using the Ensembl Variant Effect Predictor (VEP) [27268795]. We will then determine if these mutant regulatory regions are found in the promoter regions of our selected functionally related genes. For the selected regulatory SNVs, we will generate in silico wildtype and mutant alleles, using the latest version of the human reference genome hg38, as the reference (wildtype) allele. Each pair of alleles will be scored against each CGC-TF PSSM to obtain a log-odds ratio score compared to a background of genomic nucleotide frequencies. Only scores passing the CGC-TF-specific threshold and where the wildtype or mutant log odds score over background is ≥ 2 will be retained. The resulting distribution will reveal pairs of wildtype-mutant alleles from whole genome-sequenced samples with potential gain-of-binding site events, in which the mutant allele score is higher than the wildtype allele score for a particular CGC-TF. Potential loss-of-binding events will be obtained through wildtype allele score greater than the mutant allele score.

Additionally, we will use the MATCH [cit] algorithm and TRANSFAC [cit] to compare the mutant allele and wildtype allele to identify all potential TF-mutant gene interactions. For this analysis, the Yes-set (the query sequences) will be the mutant alleles, while the no-set will be the wildtype alleles. The results will provide the TFs that more frequently bind to the mutant alleles rather than the wildtype alleles. This search will not be limited to CGC-TFs.