**Instructions:**

Please take no more than 2 hours (need not be consecutive time) to do this analysis. Feel free to use any resource, except other humans. For any coding, use any programming language and/or analysis system you are comfortable with. If you can do the analysis without any coding (which is fine), please also enclose a non-proprietary sample code of your so I can see your coding style.

There is no one right answer. I am more interested in how you think and go about analysis.

Many thanks!

**The data:**

*From various sources:*

1. immune\_genes.txt. This is a list of immune-related genes.

*TCGA - Note that the below (except dlbc\_log2cna.txt) were opened and edited using Excel*

*and saved as TSV files. The identifiers, such as TCGA-FA-8693-01, represent anonymized patient IDs.*

1. dlbc\_cna.txt. -2 = homozygous deletion, -1 = heterozygous deletion, 0 = diploid, 1 = gain, and 2 = amplification of a gene.
2. dlbc\_log2cna.txt. This is log2(copy\_number + 1).
3. dlbc\_expr.txt. This is median RNA-seq gene expression.

*Broad's DepMap:*

1. broad\_info.csv - information about the Broad cell lines. The first column is the Broad ID of the cell line and the fourth column, e.g. NIHOVCAR3\_OVARY, is the name of the cell line and the site.
2. broad\_crispr.txt - a TSV matrix of cell lines (see broad\_info.csv) versus genes. The entries are the probability of negatively affecting proliferation if the gene is knocked out. I filtered this to only the genes in immune\_genes.txt for brevity.
   1. A cell line is called "dependent" on a gene if knocking out the gene negatively affects proliferation.
   2. This may be provided in four separate files, broad\_crispr.part1.txt.gz through broad\_crispr.part4.txt.gz, if emailing of large files is an issue. You will have to concatenate them into broad\_crispr.txt.

**Analysis questions:**

1. Find the genes that are most commonly homozygous deleted. For brevity, call this set of genes homDel.
   1. Part of this will require one to statistically define "commonly" and justify the definition.
2. Compare the calls in cna.txt with log2cna.txt and expr.txt for set homDel. Are they correlated?
   1. Either way, please briefly explain what is going on. If anything does not correlate, what do we do about it? This is a critical question!
      1. Take as given that there are no fusions nor deleterious mutations detected in the homDel set.
3. Let us consider only the samples where at least one gene in set homDel is homozygous deleted by cna.txt. Call this set of samples patientDel. Let us also restrict ourselves to the set of genes in immune\_genes.txt.
   1. Thus, filter dlbc\_expr.txt to only homDel samples and genes in immune\_genes.txt.
      1. Assume FGF13 should be expressed in all samples. You will need to use this assumption.
   2. Please show the genes that are very similarly expressed and very differentially expressed across the samples.
      1. You will first have to statistically define what is meant by "very similarly expressed" and "very differentially expressed".
4. For the genes in homDel that are also present in broad\_crispr.txt, determine which cell lines - if any - are dependent on at least one of the genes in homDel.
   1. Please make sure to justify, if necessary, any cutoffs you determine.