6.881 Lab 3

In the Lab portion of this class you will be looking at the distributions of the variants in the ExAC data set to analyze patterns in the human polymorphisms and study the effect of expression on the frequency of these mutations.

In order to make the lab more feasible and minimize computational power required we have provided you with a subset of the data on which to do your analysis. The data can be found in ExAC_data.txt. If you are using R the code below may help guide you through the process of making the necessary plots for analysis.

1. (a-c) First compare the expected and observed number of mutations of each type (synonymous, missense, and loss of function). The code below will guide you through loading the file and making one of the plots.

```
ExAC_data <–read.delim("ExAC_data.txt", header=T)
plot(ExAC_data$exp_syn, ExAC_data$n_syn, pch=20, xlim=c(0,1500),
ylim=c(0,1500), col="grey60", main="Observed synonynmous variants vs.
expected")
abline(0,1,lwd=3)
rs = round(cor(ExAC_data$n_syn,ExAC_data$exp_syn),digits=4)
legend("topleft", legend=bquote( R^2 == .(rs), ), bty="n")
```

(d-f)
Observe the general trend you notice relating to the deviation for the x=y line. How are the three types of mutations different? Consider the features of the genes that go into estimating the expected number of mutations. Why are there less LoF mutations expected than missense? Finally consider a scenario where the data had an even larger skew away from the x=y line. What does the degree of skew imply about the interactions between human genes and the degree of constraint on them? Given what you know about the genome sizes/features of plants and yeast what would you expect a similar plot to look like for those organisms?

2. Next, come up with your own metric, or way to identify major deviations from the y=x line (can do this for LoF or missense).

(a) Create a plot with this data highlighted. For example, you could look at difference between observed vs. expected, but to take into account gene size (larger genes with more observed & expected have greater absolute deviation)

To plot this you could add another column to your data. Let's call it pLI.

```
plot_color <–rep("grey60",length(ExAC_data$exp_lof))
plot_color[ExAC_data$pLI>0.95]<–"red"
```

```
plot(ExAC_data$exp_lof, ExAC_data$n_lof, pch=20, xlim=c(0,350), ylim=c(0,350),
col=plot_color, main="Observed synonynmous variants vs. expected")
abline(0,1,lwd=3)
```

(b) What kind of genes do you expect to see here? Explain your reasoning.

3. Subset the list by expression data — open-ended – you can consider different criteria — e.g. highly-expressed genes, or cell-type-specific genes.

(a) Using your subset of genes, does the fraction of genes classified as under constraint (highlighted in red above) change?

(b) Try at least 3 subsets. Of the ones you tried, what subset gives you the highest function of constrained genes?

(c) Use the most constrained set of genes and look at GO terms associated with these genes. Create and explain a hypothesis for what kind of gene categories are in this enriched set. Does the result differ from what you expected?