**Notes for avian missing gene analysis:**

Nathan Lazar ([nathan.lazar@gmail.com](mailto:nathan.lazar@gmail.com))

* The full script to do all the steps below is “Missing\_gene\_analysis.R”. I saved a number of the tabs from the original excel files as tab delimited text files so that they would be easier to import to R. I have included all files that the R script uses as well as .png files of the images below. I can easily generate pdfs or other formats if you would prefer. The R script should stand alone since I didn’t use any non-standard packages.
* First I checked for duplicates in the 1 to 1 human to anole ortholog table. There were two:
  + RPP14 is listed twice with two different human gene ids. Both have the same start but different endings. The anole orthologs have different starts but are close and fall in order in the gene list. Since these two entries would be treated at a block of two genes, I removed the second entry from the table before doing the permutation analysis.
    - Human loci:
      * chr3:58291974-58305816
      * chr3:58291974-58305920
    - Anole loci:
      * chr2:187318639-187323381
      * chr2:187318070-187318486)
  + LEKR1 is similarly listed twice with two human gene ids. The loci are different but close and fall in order in the gene list, so I removed the second entry.
    - Human loci:
      * chr3:156543270-156697347
      * chr3:156544096-156763918
    - Anole loci:
      * Chr3:15191770-15221921
      * Chr3:15126986-15152743
* Next I generated 1,000 lists of unique genes with the same chromosomal distributions and block sizes as were observed in the genes missing in the bird genome. Each iteration does the following:
  + For each block of genes size n in the missing gene set and I randomly choose a ‘seed’ gene on the given genome. I check that the n genes following this gene are on the same genome in both human and anole and that none of the genes are already on the list of genes for that permutation. This continues until I get a full set of 274 randomly selected genes.
* Using this list of 1,000 gene sets, I counted the number of OMIM ids associated with each gene and the number of OMIM disease terms associated with each set of genes. The distributions are shown below. The genes missing in the bird genome do not have a significantly different number of OMIM ids associated with them (two-sided permutation test p-value: 0.42). However, none of the permutations had a number of OMIM disease terms as extreme as was observed for the genes missing in the bird genome. Therefore a two-sided permutation test p-value is less than .001 indicating that the genes missing in the bird genome have a significantly fewer OMIM disease terms than would be expect by chance when accounting for the chromosomal distribution and block sizes.

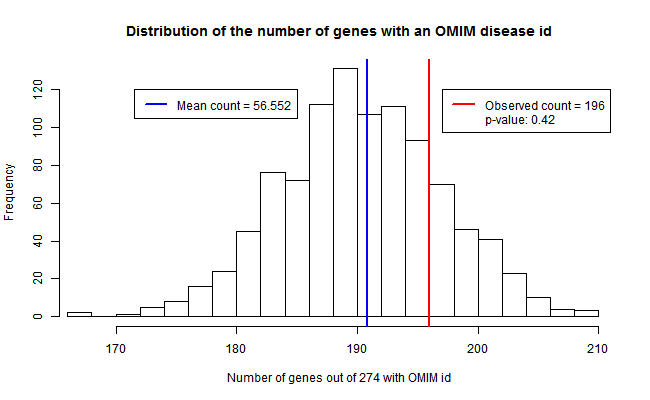


Figure 1: Genes missing in bird do not have a significantly different number of OMIM ids associated with them.

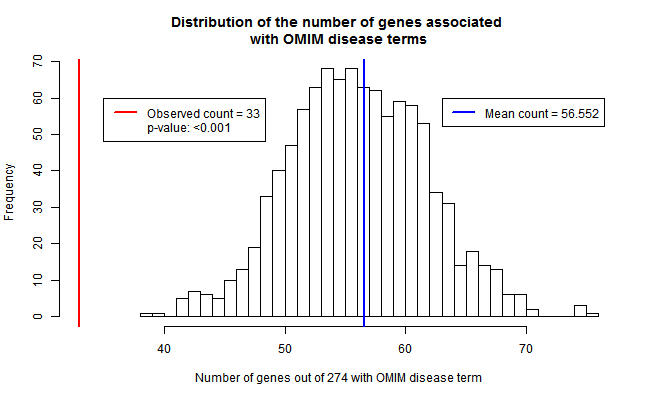
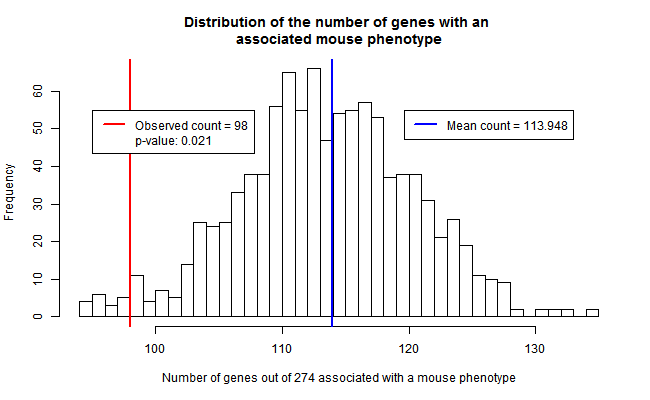


Figure 2: Genes missing in birds have significantly less OMIM disease terms than would be expected by chance.

* I then used the same list of 1,000 random gene sets to look at the number of genes missing in birds that are associated with mouse phenotypes. Of the 274 missing genes, 98 are associated with some mouse phenotype. Note that the phenotype “no abnormal phenotype detected” is not counted in this number. This is significantly less than the number of genes associated with a mouse phenotype in the permutation data (two-sided permutation test p-value: 0.021)



* Lastly I looked at the number of genes associated with each mouse phenotype in both the permutation gene sets and the set of genes missing in birds.
  + Note that the file associating genes with mouse phenotypes contains copies of pairs of gene-phenotypes if these relationships have been observed in more than one strain of mice. I believe that these instances caused double counting in the previous version of this analysis for the missing genes. For example table S10 of “HumanLizard\_e71\_1to1\_missinggenesremoved.xlsx” there are 38 missing genes with the mouse phenotype MP:0002169 “no abnormal phenotype detected”. If one counts the number of genes in the gene to phenotype table that are also in the missing gene set you get 38, but 4 of these are duplicates. If you instead count the number of missing genes (these are all distinct) that have this phenotype there are 34.
* For each phenotype I compared the number of genes associated with that phenotype in the missing gene set to the distribution of the number of genes associated with that phenotype in the permutation gene sets. I report the number of associated genes in the missing set, the mean and standard deviation for the number of associated genes in the permutation gene sets and a two-sided permutation p-value. This value is calculated as above, namely the number of permutations finding a count as extreme as that observed in the missing gene set is divided by the number of permutations. These values are in the file: Mouse\_pheno\_perm.xlsx. They seem to agree well with the results that you obtained earlier.
* I tried several approaches for multiple test correction of the above 7,874 phenotype permutation tests. For a given phenotype, if we never observed a gene count as extreme as that seen in the missing gene set then the calculated p-value is 0/1000 = 0. However since we can’t know whether such an extreme value would have occurred with more permutations, for multiple test correction these values were set to 0.001, an upper bound for the true p-value.
  + A Benjamini-Hochberg false discovery rate (FDR) adjustment showed no significant phenotypes at the 0.05 level.
  + If we assume that phenotypes with p-values set to 0.001 above would behave similarly with 10,000 permutations and we set those p-values to 0.0001, FDR correction shows 47 phenotypes with q-values < 0.05
  + Alternatively we can restrict our examination to only those phenotypes with larger effect sizes. If we only look at phenotypes for which the number of expected and observed genes differs by more than one, our number of tests is reduced to 376. With 1,000 permutations we find 11 significant FDR q-values at a 0.05 level. All of these are phenotypes where the number of missing genes is greater than one and the mean number of missing genes in permutations is less than one.
  + I would suggest using the third approach and reporting both the p-value and q-value for those phenotypes where the number of observed and expected genes differs by more than one. I’ve saved these results in the table Mouse\_pheno\_perm\_q.xlsx