# Analysis of proportion of human pathogens for different phylums from JGI IMG database

*Miroshnikova Anastasia*

*8 August 2017*

## Contents

### Introduction

The study to be published here is part of a student project which took place at Bioinformatics Institute Summer School - 2017.The data set used in the study is downloaded from JGI IMG database[1] with multiple filters. The original study was based on Deneke and Renard (2017).

The original data set contains 9563 records about sequenced bacterial genomes, but only 1858 possess information about organism phenotype. The analyzed bacteria were presented by 368 species, which belonged to one of 77 families. These families belonged to one of theese phylums:

```
##  [1] "Firmicutes"      "Proteobacteria"  "Actinobacteria"
##  [4] "Tenericutes"     "Chlamydiae"      "Bacteroidetes"
##  [7] "Fusobacteria"    "Spirochaetes"    "Synergistetes"
## [10] "Cyanobacteria"   "Deferribacteres"
```

### Hypothesis

We were interested in obtaining reliable and self-consistent sequencing data, so we decided to find out whether there is a statistically significant difference in proportion of sequenced pathogens in different phylums of bacteria. For these purpose we used **R** (version 3.4.1) and two appropriate statistical tests called *Pearson's Chi-square*:

$$X^2 = \sum \frac{(O-E)^2}{E}$$

and *Fisher's exact test*:

$$P_{cutoff} = \frac{(R_1!R_2!...R_m!)(C_1!C_2!...C_n!)}{N!\prod_{ij} a_{ij}}$$

Our zero hypothesis is that proportion of sequenced pathogens over total sequenced samples of the phylum is the same for all phylums.

The workflow is as follows:

1. Read file
2. Prepare data for analysis

[1] Here is the direct link to data download.

3. Run chi-square test and Fisher's exact test for all appropriate phylums and obtain p-value
4. Support the evidence with
   - a table of results;
   - a plot that could help demonstrate them.

## Analysis

Because of database format, there are 71 different types of phenotype. That's why before the analysis we added a new variable with only two phenotype levels - either "Pathogen" or "Non-Pathogen".

It turned out that for some phylums there was not enough observations even for Fisher's exact test (less than 8 for whole phylum). These phylums, i.e. Cyanobacteria, Deferribacteres, Synergistetes were excluded from the analysis. For the other phylums there was no data on sequenced non-pathogens, so they were also excluded from analysis (Chlamydiae, Spirochaetes, Tenericutes). The final data to be analyzed by Fisher's exact test were as follows:

```
##
##                     Non-Pathogen Pathogen
##   Actinobacteria              58      107
##   Bacteroidetes               22       29
##   Firmicutes                  86      740
##   Fusobacteria                 3       10
##   Proteobacteria             146      554
```

For chi-sqared test the phylum Fusobacteria was excluded.

So structure of our data was as follows:

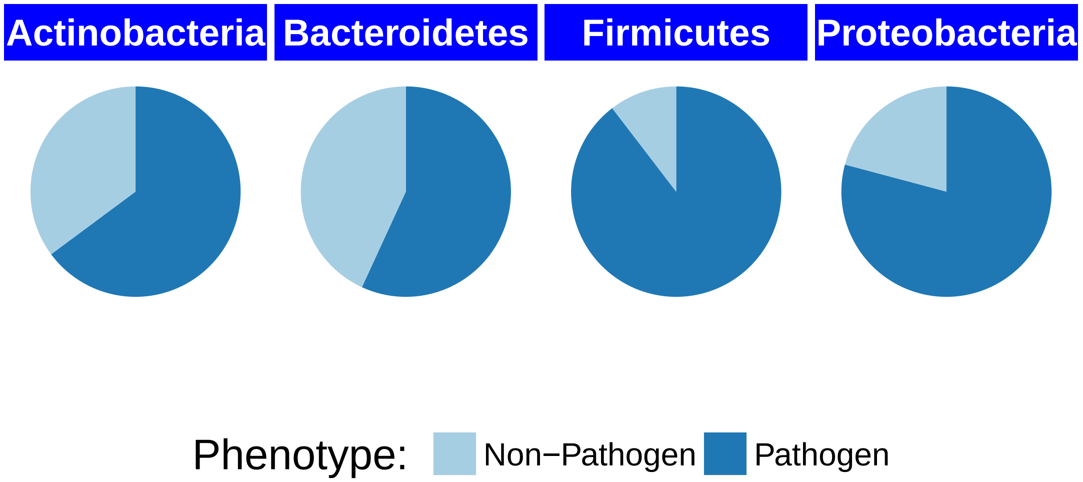

Figure 1: Relative distribution of phylums in pathogenic and non-pahtogenic sequenced samples

The p-value for the tests were $3.48 \times 10^{-18}$ for Fisher's exact test and $1.234 \times 10^{-19}$ for chi-squared test ($1.127 \times 10^{-18}$ for the same data). This let us to say that proportion of pathogens is different for at least one phylum.

For chi-squared test a matrix of residuals was

```
##
##                     Non-Pathogen Pathogen
##   Actinobacteria            5.23    -2.44
##   Bacteroidetes             4.26    -1.99
##   Firmicutes              -5.09     2.38
##   Proteobacteria           1.84    -0.86
```

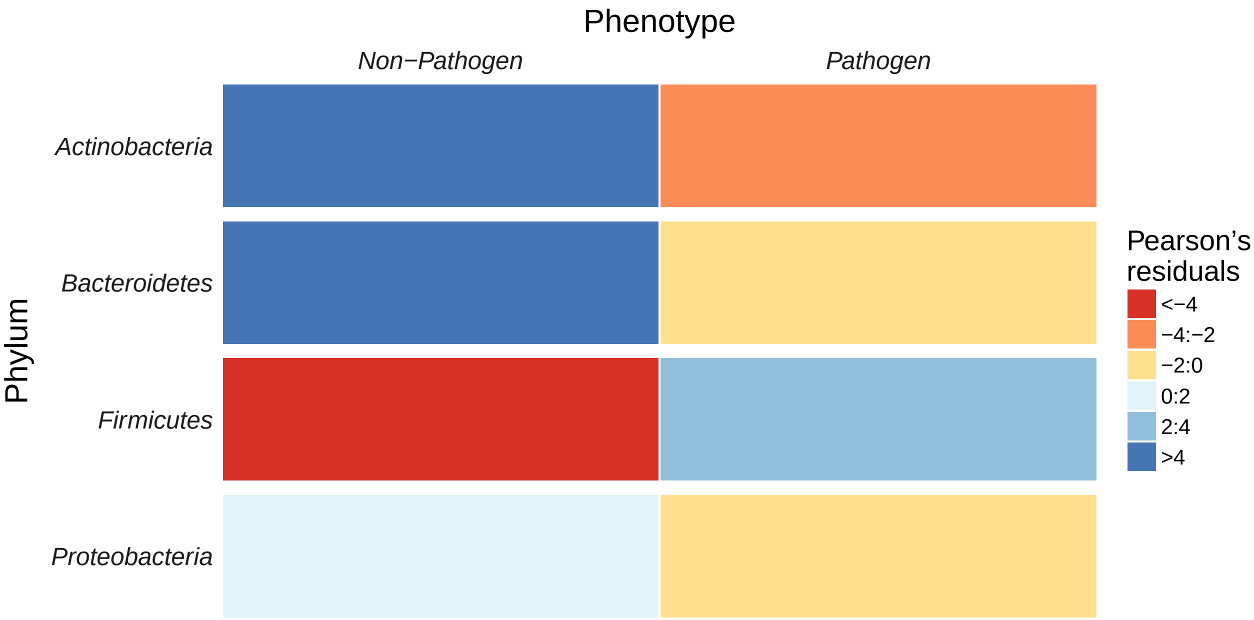

Figure 2: Pearson's residuals for different phylums

As we see, the proportion of sequenced non-pathogenic members of Actinobacteria and Bacteroidetes phylums is far more than expected, and proportion of non-pathogenic members of Firmicutes phylum is less than expected.

## Conclusion

Our hypothesis of non-uniform distribution of sequenced pathogens over different phylums in the data base of interest was proven to be valid.

## Acknowledgements

Author is very grateful to the Organizing Commitee of **Bioinformatics Institute Summer School - 2017** for letting a chance to obtain such interesting data.

## Bibliography

Deneke, Rentzsch, C., and B.Y. Renard. 2017. "PaPrBaG: A Machine Learning Approach for the Detection of Novel Pathogens from Ngs Data." *Nature Scientific Reports* 7 (39194). doi:10.1038/srep39194.