# ORFanDB: a Database of Orphan Genes in Pathogenic Bacteria

Sarah Entwistle, Xueqiong Li, Yanbin Yin (Presenter, sentwistle@niu.edu)

Department of Biological Sciences, College of Liberal Arts & Sciences, Northern Illinois University

Northern Illinois University

## Summary

ORFans are protein encoding genes with no similar genes found in taxonomically related species. Many of these unique genes are found in different types of sequence features, such as pathogenic islands (PAIs) or prophages, and may contribute to the pathogenicity of these genomes (**ref 1, 3**). **A comparison of pathogenic (P) and non-pathogenic (NP) genomes in a single species can contribute to a better understanding of the relationship between ORFans and pathogenicity.**

ORFanDB is a database containing comparative pan-genomic information of ORFans for pathogenic and non-pathogenic bacterial species. Utilizing publicly available data as well as a new ORFan classification scheme, the results of an analysis of ORFans in different sequence features are contained in the database. ORFanDB will also allow users to classify and annotate their own ORFans using tools developed in the Yin lab (**ref 2**).

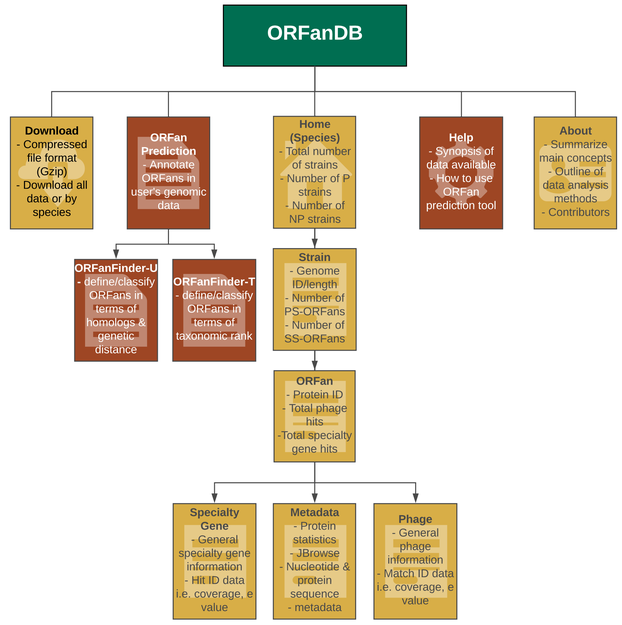
## Methods

- Nine genera of interest identified (**ref 3**)
  - At least 5 pathogenic genomes
  - At least 5 non-pathogenic genomes
- DIAMOND – identify ORFans
  - e-value < 0.01
- PHASTER – identify ORFans found in prophages
- IslandViewer – identify ORFans found in pathogenic islands.
- R
  - Statistical analysis (Wilcoxon test)
  - Data visualization

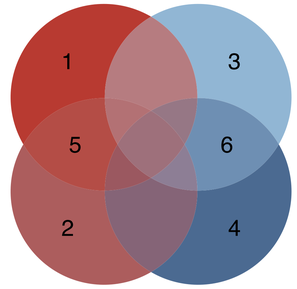
## Genera Statistics

| Genus | Phylum | Total genomes | Pathogenic genomes (P) | Non-pathogenic genomes (NP) |
|---|---|---|---|---|
| *Bacillus* | *Firmicutes* | 79 | 34 | 45 |
| *Burkholderia* | *Proteobacteria* | 33 | 27 | 6 |
| *Clostridium* | *Firmicutes* | 32 | 17 | 15 |
| *Corynebacterium* | *Actinobacteria* | 51 | 35 | 16 |
| *Escherichia* | *Proteobacteria* | 57 | 47 | 10 |
| *Listeria* | *Firmicutes* | 40 | 28 | 12 |
| *Mycobacterium* | *Actinobacteria* | 54 | 44 | 10 |
| *Pseudomonas* | *Proteobacteria* | 51 | 18 | 33 |
| *Streptococcus* | *Firmicutes* | 108 | 90 | 18 |

## ORFanDB

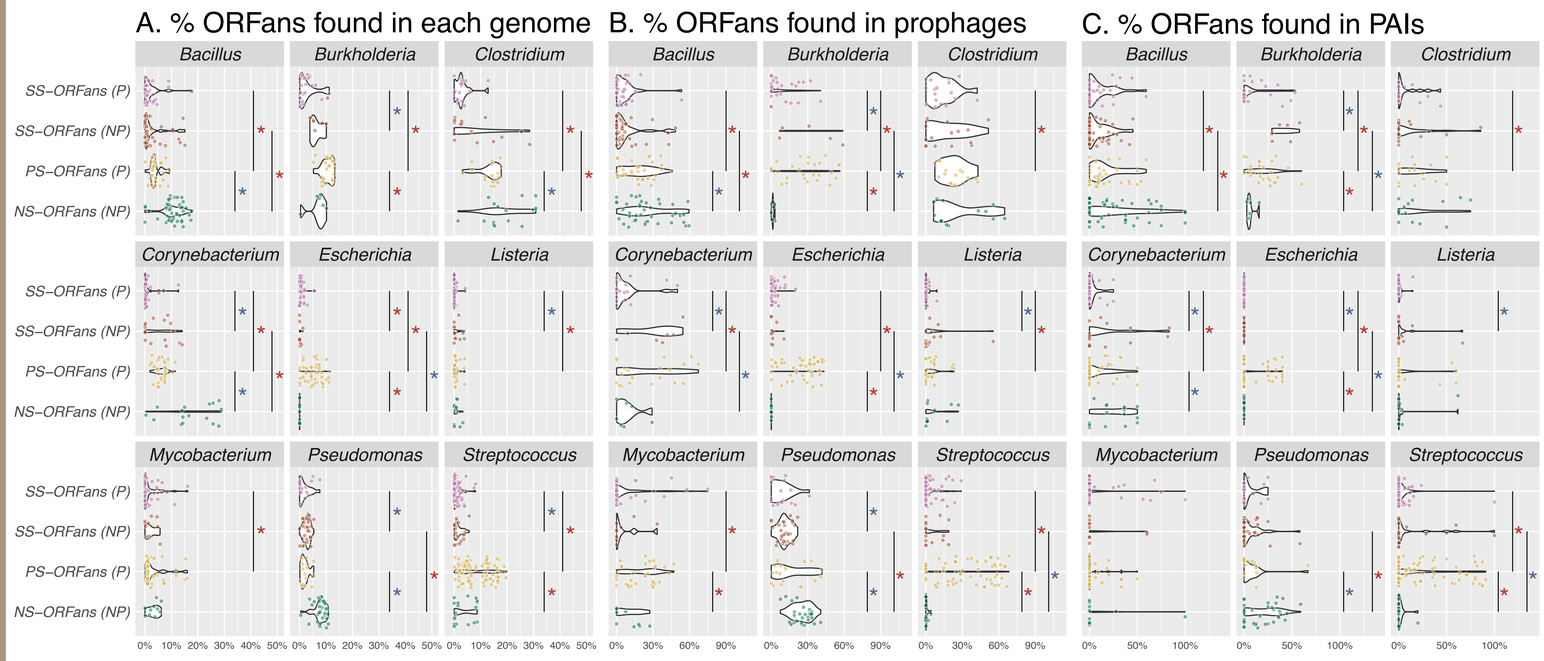



## ORFan Classification Scheme

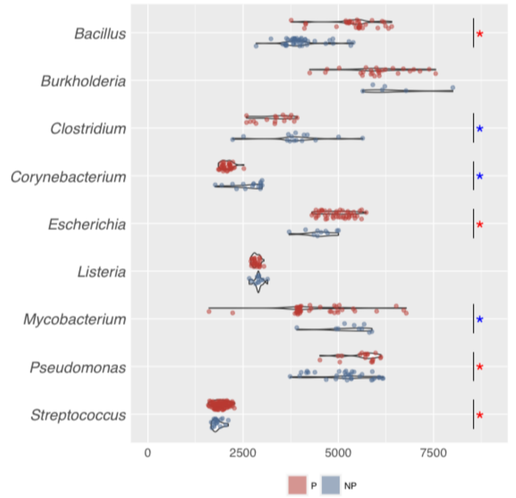



- Strain specific ORFans (SS-ORFans) – genes existing in only a single genome – 1, 2, 3, 4
- Pathogen specific ORFans (PS-ORFans) – genes existing in two or more pathogenic genomes but not in non-pathogenic genomes - 5
- Non-pathogen specific ORFans (NS-ORFans) – genes existing in two or more non-pathogenic genomes but not in pathogenic genomes – 6
- Pathogenic genomes (P) – Red circles
- Non-pathogenic genomes (NP) – Blue circles

## Fig 1: Comparing different groups of ORFans (*: p-value < 0.05, *: p-value > 0.95)




## Fig 2: Comparing the total # of genes




*: p-value < 0.05 (support P > NP), *: p-value > 0.95 (support P < NP)

## Funding

NIH National Institutes of Health

Northern Illinois University

## Conclusion

- We hypothesized that (i) P genomes are larger than NP genomes (Fig 2), (ii) P genomes have more ORFans than NP genomes (Fig 1A), (iii) P genomes have more ORFans located in prophages (Fig 1B) and PAIs (Fig 1C) than NP genomes, and (iv) there should be more PS-ORFans than SS-ORFans (Fig 1A), and more PS-ORFans located in prophages and PAIs than SS-ORFans in P genomes (Fig 1B and 1C). Contrary to previous research, all these hypotheses except for the last one were not well supported in our comparative study.
- Overall we found that the association between ORFans and pathogenesis is at best weak and varies strongly between different bacterial genera. For example, the most characterized *Escherichia* genus accepts our hypothesis in most comparisons. However, the *Corynebacterium and Pseudomonas* genera reject our hypothesis in most comparisons, suggesting that ORFans may not be the major factor in their pathogenicity.
- Indeed, other than gene gain through horizontal gene transfer from phages or other bacteria, there are other important factors that can also lead to pathogenesis such as gene loss due to genome reduction, modifying core genome such as SNPs, indels and recombinations.

## References

1. Busby, B., Kristensen, D. M., & Koonin, E. V. (2012). Contribution of phage-derived genomic islands to the virulence of facultative bacterial pathogens. *Environmental Microbiology, 15*(2), 307-312. doi:10.1111/j.1462-2920.2012.02886.x
2. Ekstrom, A., & Yin, Y. (2016). ORFanFinder: Automated identification of taxonomically restricted orphan genes. *Bioinformatics, 32*(13), 2053-2055. doi:10.1093/bioinformatics/btw122
3. Sui, S. J., Fedynak, A., Hsiao, W. W., Langille, M. G., & Brinkman, F. S. (2009). The Association of Virulence Factors with Genomic Islands. *PLoS ONE, 4*(12). doi:10.1371/journal.pone.0008094

## ORFan Analysis Observations

| $H_0$ | SS-ORFans (P) > SS-ORFans (NP) | | | PS-ORFans (P) > NS-ORFans (NP) | | | PS-ORFans (P) > SS-ORFans (P) | | | NS-ORFans (NP) > SS-ORFans (NP) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Figure | A | B | C | A | B | C | A | B | C | A | B | C |
| Accept | 1 | 0 | 0 | 3 | 4 | 3 | 8 | 8 | 6 | 4 | 2 | 2 |
| Reject | 5 | 4 | 4 | 4 | 2 | 2 | 0 | 0 | 0 | 1 | 4 | 3 |
| Not Significant | 3 | 5 | 5 | 2 | 3 | 4 | 1 | 1 | 3 | 4 | 3 | 4 |