Supervised Machine Learning Approach for Prediction of *Legionella pneumophila* Serogroups from Whole Genome Sequencing Data

Shatavia S. Morrison, Natalia A. Kozak-Muiznieks, Jason A. Caravas, Brian H. Raphael, and Jonas M. Winchell

Pneumonia Response and Surveillance Laboratory, Respiratory Diseases Branch, Centers for Disease Control and Prevention, Atlanta, GA

Background

The majority of Legionnaire's Disease (LD) cases are due to a single species, *Legionella pneumophila*. This species consists of at least 17 serogroups, with serogroup 1 (Lp1) frequently isolated from clinical and environmental samples (70- 92%). Since the 1980s there have been several reports of non-Lp1 isolates linked to LD outbreak investigations and clinical cases. Even though non-Lp1 cases have been reported along the same historical timeframe as Lp1, Lp1 is the only serogroup to date that has both molecular and genetic based detection assays. Current genetic techniques classify *L. pneumophila* isolates into either Lp1 and non-Lp1 categories, without the possibility of further characterization unless wet bench work is performed. With bacterial sequences more readily available, there is an urgent need to develop sequence-based tools for identification of all *L. pneumophila* serogroup. We developed a supervised machine learning model to predict *L. pneumophila* serogroup classification from short read sequencing data targeting the lipopolysaccharide biosynthesis region within the genome.

Results

Data Set 1: Pneumonia Response and Surveillance Laboratory

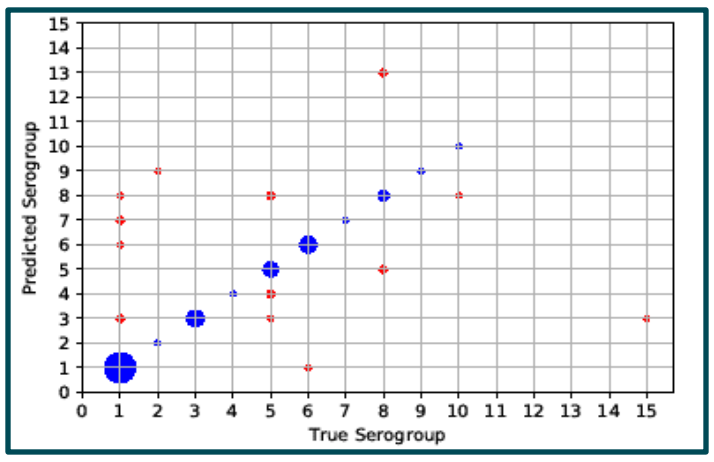


Figure 1. Comparison of Predicted Serogroup and True Serogroup (Wet Bench Validated) Classification for Data Set 1. 🔵 = Correct prediction and 🔴 = Incorrect prediction. The size of the circle plot is proportional to the number of isolates classified for each serogroup.

We sequenced 473 *L. pneumophila* isolates not previously seen by the model to test model robustness. Overall, our model achieved a prediction accuracy of 95.9% for this data set.

Table 1. Summary of Prediction Accuracy Per Serogroup for Data Set 1.

Serogroup	# of Isolates	# of Correct Prediction	# of Incorrect Prediction	Success Rate	Error Rate
1	427	420	7	98.36%	1.64%
2	2	1	1	50.00%	50.00%
3	9	9	0	100.00%	0.00%
4	1	1	0	100.00%	0.00%
5	11	7	4	63.64%	36.36%
6	10	9	1	90.00%	10.00%
7	1	1	0	100.00%	0.00%
8	8	4	4	50.00%	50.00%
9	1	1	0	100.00%	0.00%
10	2	1	1	50.00%	50.00%
15	1	0	1	0.00%	100.00%

* Lp11 – Lp14,Lp16,Lp17 were not represented in this dataset. * Success rate is defined as the total number of isolates which prediction matches the isolate's recorded direct fluorescent antibody (DFA). Error rate is defined as the total number of predictions that did not match the isolate's recorded DFA.

Data Set 2: *Legionella* Laboratory External Collaborator

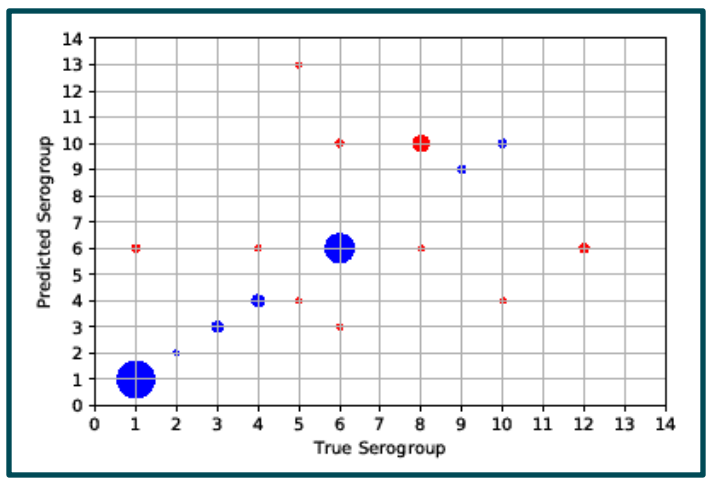


Figure 2. Comparison of Predicted Serogroup and True Serogroup (Wet Bench Validated) Classification for Data Set 2. 🔵 = Correct prediction and 🔴 = Incorrect prediction. The size of the circle plot is proportional to the number of isolates classified for each serogroup.

This data set consisted of 109 *L. pneumophila* isolates. The purpose of this data set was to test our model to see if it can overcome any potential sequencing bias in data that was not generated in our lab. Our model achieved a prediction accuracy of 80.7% for this data set.

Table 2. Summary of Prediction Accuracy Per Serogroup for Data Set 2.

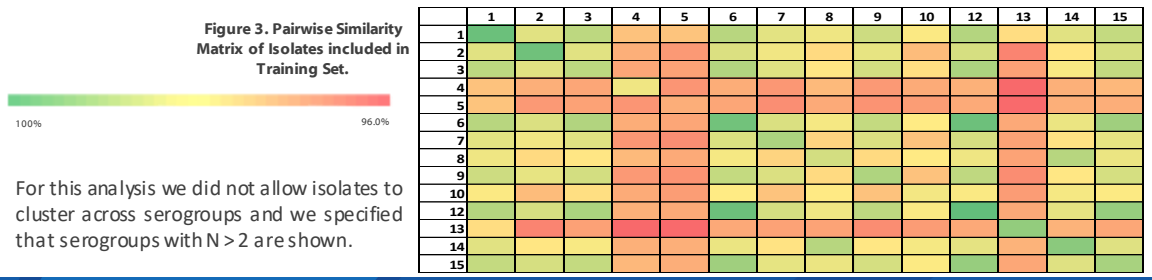
Serogroup	# of Isolates	# of Correct Prediction	# of Incorrect Prediction	Success Rate	Error Rate
1	48	46	2	95.83%	4.17%
2	1	1	0	100.00%	0.00%
3	4	4	0	100.00%	0.00%
4	6	5	1	83.33%	16.67%
5	2	0	2	0.00%	100.00%
6	31	28	3	90.32%	9.68%
8	9	0	9	0.00%	100.00%
9	2	2	0	100.00%	0.00%
10	3	2	1	66.67%	33.33%
12	3	0	3	0.00%	100.00%

* Lp7, Lp11, Lp13, Lp14 -Lp17 were not represented in this dataset. * Success rate is defined as the total number of isolates which prediction matches the isolate's recorded direct fluorescent antibody (DFA). Error rate is defined as the total number of predictions that did not match the isolate's recorded DFA.

Comparative Sequence Analysis of 181 *L. pneumophila* Serogroup LPS Region

We investigated the LPS region by performing a pairwise sequence similarity comparison. We calculated the average distance between all isolates associated with a specific serogroup. The average distance was used to represent the serogroup in the pairwise analysis when compared to other serogroups.

Figure 3. Pairwise Similarity Matrix of Isolates included in Training Set.

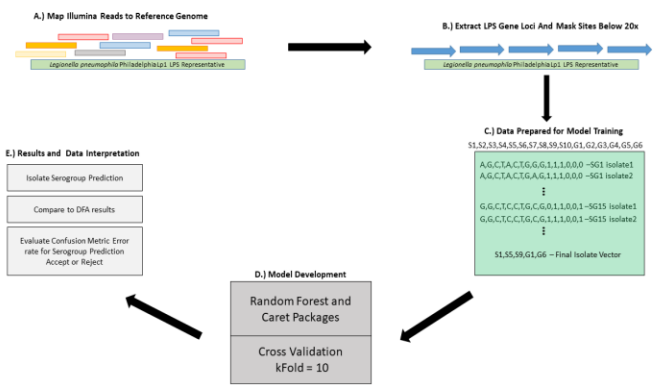


For this analysis we did not allow isolates to cluster across serogroups and we specified that serogroups with N > 2 are shown.

Lp6 and Lp12 shared a high sequence similarity (99.67%), this maybe one of several factors that impacts our ability to predict Lp12.

Lp5 isolates have a similarity range between 3-4.5%, compared to 1-2% observed in other serogroups, leading us to hypothesize that there maybe distinct lineages within Lp5.

Methods and Materials



A.) 181 *Legionella pneumophila* isolates representing various serogroups geographical locations, and source types (clinical and environmental) were selected to train the Random Forest model.

B.) Several studies have reported that the lipopolysaccharide biosynthesis region within the L. pneumophila genome is associated with serogroup classification [1-4]. This was the target region for this analysis.

C.) SNPs were identified based on 25x coverage and gene presence-absence was assigned based on 80% gene coverage to reference loci

D.) H2O – AutoML algorithm was used to identified the optimal machine learning approach for classification. Based on the results, we implemented the Random Forest model in R and used Caret package for kFold cross-validation. Laplace smoothing was applied to input matrix prior to model construction.

E.) Prediction results are compared to error rates associated with the cross-validation data set

Future Work

- Convert the work presented here into a backend infrastructure for a user-friendly tool to be hosted on the OAMD Bioinformatics Portal.
- Conduct further analyses on non-Lp1 serogroups using Pacific Bioscience sequences to investigate gene arrangement.
- Refine prediction model to consider gene arrangement to increase prediction accuracy for current poor performers.
- May revise current schema to consider sub-groups within specific serogroups
- Improvement informatics workflow to decrease data-to-result turn around time.

Conclusions

- We constructed a successful machine learning model for the prediction of *Legionella pneumophila* serogroup information from shot-gun short read data.
- We achieved a prediction accuracy > 80% with our model design and implementation.
- To our knowledge this work is the first of its kind to address non-Lp1 genetic serogroup typing.
- There may genetic explanations as to why specific serogroups (Lp5,Lp8, Lp6, Lp10, and Lp12) are poor performers with our current model's approach.
- In this dataset Lp5 has the highest range of genetic diversity, which may make it difficulty for our approach in identifying commonalities to structure a Lp5 specific DNA pattern. There maybe sub-groups we may need to investigate to improve the classification serogroup schema.
- Lp6 and Lp12 are nearly identical (99%) from a genetic perspective. We are able to predict Lp6 with reliability compared to Lp12. This may be due to the "small n" representation of Lp12 during model training.

References:

1.Cao B, Tian Z, Wang S, Zhu Z, Sun Y, Feng L, Wang L: **Structural comparison of O-antigen gene clusters of Legionella pneumophila and its application of a serogroup-specific multiplex PCR assay.** *Antonie Van Leeuwenhoek* 2015, **108**(6):1405-1423.
2.Cazalet C, Jarraud S, Ghavi-Helm Y, Kunst F, Glaser P, Etienne J, Buchrieser C: **Multigenome analysis identifies a worldwide distributed epidemic Legionella pneumophila clone that emerged within a highly diverse species.** *Genome Res* 2008, **18**(3):431-441.
3.Helbig JH, Kurtz JB, Pastoris MC, Pelaz C, Luck PC: **Antigenic lipopolysaccharide components of Legionella pneumophila recognized by monoclonal antibodies: possibilities and limitations for division of the species into serogroups.** *J Clin Microbiol* 1997, **35**(11):2841-2845.
4.Petzold M, Thurner A, Menzel S, Mouton JW, Heuner K, Luck C: **A structural comparison of lipopolysaccharide biosynthesis loci of Legionella pneumophila serogroup 1 strains.** *BMC Microbiol* 2013, **13**:198.

Acknowledgements:

- CDC Advanced Molecular Detection Initiative : Awarded Project Titled, "Whole Genome Sequence Analysis Tools for Rapidly Investigating *Legionella* outbreak"
- Office of Advanced Molecular Detection: Scientific Computing and Bioinformatics Support
- Arizona Public State Health Department : *Legionella* Group

Contact Information:
Shatavia Morrison, PhD - SMorrison@cdc.gov
Brian Raphael, PhD - BRaphael@cdc.gov
Jonas Winchell, PhD - JWinchell@cdc.gov

Poster # 21

National Center for Immunization and Respiratory Diseases

Division of Bacterial Diseases