The Journal of Open Source Software

SeroTools: a Python package for *Salmonella* serotype data analysis

Joseph D. Baugher, Ph.D.[1]

1 Center for Food Safety and Applied Nutrition, U.S. Food and Drug Administration

DOI: 10.21105/joss.02556

Software

- Review
- Repository
- Archive

Editor: Mark A. Jensen

Reviewers:

- @amoeba
- @Maghnuso

Submitted: 20 July 2020
Published: 05 September 2020

Summary

Subtyping, the ability to differentiate and characterize closely related microorganisms, has historically been a critical component of successful outbreak identification and traceback efforts employed by public health researchers and regulatory agencies for foodborne pathogens. Serological subtyping (or serotyping) has been the standard approach, largely based on antibody binding to surface antigens (Henriksen, 1978). The identification of specific antigenic factors has facilitated the creation of serotyping schemes, which define each serovar using a specific (generally unique) combination of antigenic factors. Serotyping schemes have been developed to assist in characterization of many microorganisms, including pathogens such as *Salmonella*, *E. coli*, *Shigella* (Strockbine, Bopp, Fields, Kaper, & Nataro, 2015), *Streptococcus* (Spellerberg & Brandt, 2015), and *H. influenzae* (Ledeboer & Doern, 2015).

Salmonella is a major foodborne pathogen for which serotyping has played a fundamental monitoring role for over 50 years(CDC, n.d.). *Salmonella* serotyping is generally based on antibody binding to the O antigen (a surface antigen) and one or more H antigen phases (flagellar antigens) (Andrews, Wang, Jacobson, & Hammack, 2019; Strockbine et al., 2015). The White-Kauffmann-Le Minor (WKL) *Salmonella* scheme specifies the naming and formatting conventions for *Salmonella* serotyping data and the antigenic factors (and other characteristics) which define each serovar (Grimont & Weill, 2007). SeroTools includes the 2007 WKL scheme (Grimont & Weill, 2007) and updates (Bugarel et al., 2015; Guibourdenche et al., 2010; Issenhuth-Jeanjean et al., 2014).

The WKL scheme currently recognizes two species of *Salmonella*, *S. enterica* and *S. bongori*. *S. enterica* is comprised of six subspecies (subsp.): *enterica* (I), *salamae* (II), *arizonae* (IIIa), *diarizonae* (IIIb), *houtenae* (IV) and *indica* (VI). Note that *S. bongori* is still frequently designated as subsp. V for scheme consistency, although it is no longer considered a subspecies of *S. enterica*. The WKL scheme assigns a unique name (e.g. serovar Enteritidis) to each of the serovars of *S. enterica* subsp. *enterica* (I), while the serovars representing the other subspecies are referred to by their antigenic formulae. The antigenic formula formatting is defined by the WKL scheme and is demonstrated for serovar Agona in Figure 1. The formula contains a subspecies designation and a colon-separated list of antigenic factors for which the following fields are required: O antigen, phase 1 H antigen, and phase 2 H antigen. The field for 'Other H' antigen includes R phases and third phases and is present only when populated. An antigenic formula may include additional annotation such as:

1. *Square* brackets to indicate optional factors, (e.g. I 1,4,**[5]**,12:f,g,s:**[1,2]**:**[z27]**,**[z45]**).
2. *Underlining* to indicate O factors present only in the presence of the converting phage, represented here and in SeroTools as optional (with *square* brackets) due to the inability to capture typographical formatting in plain text, (e.g. I **[1]**,9,12:e,h:1,5).
3. *Curly* brackets to indicate mutually exclusive factors, (e.g. I 3,**{10}{[15]}**:k:1,5).
4. *Parentheses* to indicate factors which are weakly agglutinable, (e.g. IIIb **(6)**,14:k:z53).

5. A *dash* to indicate a missing antigen, (e.g. I 1,9,12:g,m:–).

These additional annotations are captured in the SeroTools repository and employed for determination of congruence between serovars.

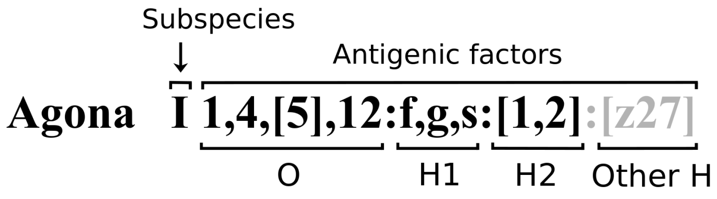

Figure 1: Standard formatting of the antigenic formula.

Statement of Need

SeroTools addresses multiple critical needs for the efficient analysis of *Salmonella* serotyping data within the public health community. In recent years, significant technological advances have resulted in a wide range of molecular-based subtyping options, including highly sensitive approaches based on whole genome sequencing (WGS). One such approach involves the application of software tools to WGS data for *in silico* serovar prediction (Joensen, Tetzschner, Iguchi, Aarestrup, & Scheutz, 2015; Laing, Bessonov, Sung, & La Rose, n.d.; Watts & Holt, 2019; Wu, Lau, Lee, Lau, & Payne, 2019; Zhang et al., 2019, 2015), including real-time prediction (Feng et al., 2020). SeqSero (a *Salmonella*-specific tool) and other *in silico* serovar designation tools have been adopted by U.S. public health agencies as an alternative to serological testing and for quality control applications (Dowdy, 2017; Timme, Sanchez Leon, & Allard, 2019). The advent of new methodologies for serovar determination has engendered a need for method-comparison studies, and has sparked a growing collection of recent publications comparing various laboratory-based and *in silico* serovar predictions (Banerji, Simon, Tille, Fruth, & Flieger, 2020; Cooper et al., 2020; Diep et al., 2019; Ibrahim & Morin, 2018; Tang et al., 2019; Yachison et al., 2017; Zhang et al., 2019, 2015). In light of the growing interest in *in silico* serovar prediction and serotyping method-comparison studies, SeroTools provides unique tools which fill multiple gaps in the analysis process. It serves as the only multiformat WKL repository accessible for software development. Currently the WKL scheme is available only as a pdf document (Grimont & Weill, 2007) and as Python lists in SeqSero (Zhang et al., 2015) and SeqSero2 (Zhang et al., 2019). SeroTools also provides the only existing tools for querying the WKL scheme, comparing serovars for congruence, and predicting the most abundant serovar for clusters of isolates.

Functionality and Features

The SeroTools Python package provides the following functionality:

1. Repository –
   - SeroTools includes an updated WKL repository in multiple formats, including Python data structures (a pandas DataFrame, dictionaries, and lists) and spreadsheets (Excel and tab-delimited). The repository includes fields representing serovar name, antigenic formula, species, subspecies, O antigen, phase 1 H

antigen, phase 2 H antigen, other H antigens, the new O group designation (e.g. O:2), and the old O group designation (e.g. A).

2. Toolkit –
   - **query** - SeroTools provides the ability to easily query the WKL repository with serovar names or antigenic formulas.
   - **compare** - SeroTools provides a convenient method for automated comparison of serovar designations, including increased differentiation for levels of congruence.
   - **cluster** - SeroTools includes methods for robust determination of the most abundant serovar for a cluster of isolates.
3. Additional functionality –
   - SeroTools includes Pythonic data structures and a host of utility functions for analyzing and manipulating large *Salmonella* serovar datasets. Other functionality includes the ability to determine the antigenic factors common to a group of serovars.

SeroTools defines four levels of congruence for use in querying the repository and comparing serovars. Note - *optional* factors as referenced below include optional, exclusive, and weakly agglutinable factors, as specified in the WKL scheme.

1. **Exact** matches must meet **one** of the following criteria:
   - The serovar designations are the identical string.
     For example:

     ```
     Corvallis                      Corvallis
     I 8,[20]:z4,z23:[z6]           I 8,[20]:z4,z23:[z6]
     ```

   - Every antigenic factor (*required* or *optional*) matches.
     For example:

     ```
     Corvallis                      I 8,[20]:z4,z23:[z6]
     I 8,[20]:z4,z23:[z6]           I 8,20:z4,z23:z6
     I 1,3,10,19:f,g,t:1,(2),7      I 1,3,10,19:f,g,t:1,2,7
     ```

   - The subspecies designations are identical and neither serovar designation includes any antigenic factors.
     For example:

     ```
     I ::                           I -:-:-
     II :                           II -:
     ```

2. **Congruent** matches must meet **all** of the following criteria:
   - The subspecies field must be present either for both serovars or for neither.
   - All *required* antigenic factors match.
   - Any differences are due to the presence/absence of *optional* factors.
     For example:

     ```
     I 6,7,14:g,m,s:-               I 6,7,[14],[54]:g,m,[p],s:-
     I 6,7:g,m,s:-                  I 6,7,[14],[54]:g,m,[p],s:[1,2,7]
     Amager var. 15+                Amager
     I 3,15:y:1,2:[z45]             I 3,{10}{15}:y:1,2:[z45]
     6,7:k:[z6]                     6,7:k:-
     ```

3. **Minimally congruent** matches must meet the following criteria:

- Every antigen of at least one serovar can be considered a formal subset of the corresponding antigen (no direct conflicts). Note - the empty set (–) is a subset of every set.
  For example:

  ```
  I 6,7,14,[54]:g,m,[p],s:-   6,7,[14],[54]:g,m,[p],s:-
  I                           I 6,7,8,[14],[54]:g,m,[p],s:-
  I 7:g:-                     I 6,7:g,m,s:-
  Gallinarum                  Enteritidis
  ```

4. **Incongruent** matches must meet the following criteria:

- Any comparison which is not at least minimally congruent.
  For example:

  ```
  I                           II
  I 1:                        1 2:
  Javiana                     Saintpaul
  I 7,8:g,m,s:-               I 6,7,[14],[54]:g,m,[p],s:[1,2,7]
  I 4,5:a,b:6,7               I 5:a,b,c:6,7
  ```

The 'minimally congruent' designation is unique to SeroTools and is useful for distinguishing between two scenarios: serovars which differ due to sample misannotation (truly incongruent) and serovars derived from correctly annotated samples with variation based solely on missing information. When comparing serovar predictions, minor differences may be expected due to method-specific irregularities, for example, reagent variation for laboratory-based techniques or sequencing read coverage for *in silico* techniques. Our assumption is that these minor method-specific differences are more likely manifested as missing data (e.g. all but one of the correct factors were detected) than direct conflicts.

Links

Documentation: https://serotools.readthedocs.io/en/latest/readme.html

Source Code: https://github.com/CFSAN-Biostatistics/serotools

PyPI Distribution: https://pypi.python.org/pypi/serotools

References

Andrews, W. H., Wang, H., Jacobson, A., & Hammack, T. S. (2019). Chapter 5: *Salmonella*. In *Bacteriological Analytical Manual*. U.S. Food and Drug Administration. Retrieved from https://www.fda.gov/food/laboratory-methods-food/bacteriological-analytical-manual-bam-chapter-5-salmonella

Banerji, S., Simon, S., Tille, A., Fruth, A., & Flieger, A. (2020). Genome-based *Salmonella* serotyping as the new gold standard. *Sci Rep*, *10*(1), 4333. doi:10.1038/s41598-020-61254-1

Bugarel, M., den Bakker, H. C., Nightingale, K. K., Brichta-Harhay, D. M., Edrington, T. S., & Loneragan, G. H. (2015). Two draft genome sequences of a new serovar of *Salmonella enterica*, serovar Lubbock. *Genome Announc*, *3*(2). doi:10.1128/genomeA.00215-15

CDC. (n.d.). Serotypes and the importance of serotyping *Salmonella*. U.S. Centers for Disease Control and Prevention. Retrieved from https://www.cdc.gov/salmonella/reportspubs/salmonella-atlas/serotyping-importance.html

Cooper, A. L., Low, A. J., Koziol, A. G., Thomas, M. C., Leclair, D., Tamber, S., Wong, A., et al. (2020). Systematic evaluation of whole genome sequence-based predictions of *Salmonella* serotype and antimicrobial resistance. *Front Microbiol*, *11*, 549. doi:10.3389/fmicb.2020.00549

Diep, B., Barretto, C., Portmann, A. C., Fournier, C., Karczmarek, A., Voets, G., Li, S., et al. (2019). *Salmonella* serotyping; Comparison of the traditional method to a microarray-based method and an *in silico* platform using whole genome sequencing data. *Front Microbiol*, *10*, 2554. doi:10.3389/fmicb.2019.02554

Dowdy, J. (2017). Deng awarded UGA creative research medal for *Salmonella* classification software. *Growing Georgia*. Retrieved from https://georgia.growingamerica.com/news/2017/05/deng-awarded-uga-creative-research-medal-salmonella-classification-software

Feng, X., Ge, C., Luo, H., Li, S., Wiedmann, M., Deng, X., Zhang, G., et al. (2020). Evaluation of real-time nanopore sequencing for *Salmonella* serotype prediction. *Food Microbiol*, *89*, 103452. doi:10.1016/j.fm.2020.103452

Grimont, P. A., & Weill, F. X. (2007). *Antigenic formulae of the Salmonella serovars* (9th ed.). Paris, France: WHO Collaborating Center for Reference and Research on *Salmonella*, Institut Pasteur. Retrieved from https://www.pasteur.fr/sites/default/files/veng_0.pdf

Guibourdenche, M., Roggentin, P., Mikoleit, M., Fields, P. I., Bockemühl, J., Grimont, P. A., & Weill, F. X. (2010). Supplement 2003-2007 (no. 47) to the White-Kauffmann-Le Minor scheme. *Res Microbiol*, *161*(1), 26–9. doi:10.1016/j.resmic.2009.10.002

Henriksen, S. D. (1978). Chapter I: Serotyping of Bacteria. In *Methods in Microbiology* (12th ed.). Academic Press. doi:10.1016/S0580-9517(08)70355-6

Ibrahim, G. M., & Morin, P. M. (2018). *Salmonella* serotyping using whole genome sequencing. *Front Microbiol*, *9*, 2993. doi:10.3389/fmicb.2018.02993

Issenhuth-Jeanjean, S., Roggentin, P., Mikoleit, M., Guibourdenche, M., Pinna, E. de, Nair, S., Fields, P. I., et al. (2014). Supplement 2008-2010 (no. 48) to the White-Kauffmann-Le Minor scheme. *Res Microbiol*, *165*(7), 526–30. doi:10.1016/j.resmic.2014.07.004

Joensen, K. G., Tetzschner, A. M., Iguchi, A., Aarestrup, F. M., & Scheutz, F. (2015). Rapid and easy *in silico* serotyping of *Escherichia coli* isolates by use of whole-genome sequencing data. *J Clin Microbiol*, *53*(8), 2410–26. doi:10.1128/JCM.00008-15

Laing, C., Bessonov, K., Sung, S., & La Rose, C. (n.d.). ECTyper. Retrieved February 27, 2020, from https://github.com/phac-nml/ecoli_serotyping

Ledeboer, N. A., & Doern, G. V. (2015). Chapter 36: *Haemophilus*. In *Manual of Clinical Microbiology* (11th ed.). Washington, DC: ASM Press. doi:10.1128/9781555817381.ch36

Spellerberg, B., & Brandt, C. (2015). Chapter 22: *Streptococcus*. In *Manual of Clinical Microbiology* (11th ed.). Washington, DC: ASM Press. doi:10.1128/9781555817381.ch22

Strockbine, N. A., Bopp, C. A., Fields, P. I., Kaper, J. B., & Nataro, J. P. (2015). Chapter 37: *Escherichia, Shigella, and Salmonella*. In *Manual of Clinical Microbiology* (11th ed.). Washington, DC: ASM Press. doi:10.1128/9781555817381.ch37

Tang, S., Orsi, R. H., Luo, H., Ge, C., Zhang, G., Baker, R. C., Stevenson, A., et al. (2019). Assessment and comparison of molecular subtyping and characterization methods for *Salmonella*. *Front Microbiol*, *10*, 1591. doi:10.3389/fmicb.2019.01591

Timme, R. E., Sanchez Leon, M., & Allard, M. W. (2019). Utilizing the public GenomeTrakr database for foodborne pathogen traceback. *Methods Mol Biol*, *1918*, 201–212. doi:10.1007/978-1-4939-9000-9_17

Watts, S. C., & Holt, K. E. (2019). Hicap: *In silico* serotyping of the *Haemophilus influenzae* capsule locus. *J Clin Microbiol*, *57*(6). doi:10.1128/JCM.00190-19

Wu, Y., Lau, H. K., Lee, T., Lau, D. K., & Payne, J. (2019). *In silico* serotyping based on whole-genome sequencing improves the accuracy of *Shigella* identification. *Appl Environ Microbiol*, *85*(7). doi:10.1128/AEM.00165-19

Yachison, C. A., Yoshida, C., Robertson, J., Nash, J. H. E., Kruczkiewicz, P., Taboada, E. N., Walker, M., et al. (2017). The validation and implications of using whole genome sequencing as a replacement for traditional serotyping for a national *Salmonella* reference laboratory. *Front Microbiol*, *8*, 1044. doi:10.3389/fmicb.2017.01044

Zhang, S., den Bakker, H. C., Li, S., Chen, J., Dinsmore, B. A., Lane, C., Lauer, A. C., et al. (2019). SeqSero2: Rapid and improved *Salmonella* serotype determination using whole-genome sequencing data. *Appl Environ Microbiol*, *85*(23). doi:10.1128/AEM.01746-19

Zhang, S., Yin, Y., Jones, M. B., Zhang, Z., Deatherage Kaiser, B. L., Dinsmore, B. A., Fitzgerald, C., et al. (2015). *Salmonella* serotype determination utilizing high-throughput genome sequencing data. *J Clin Microbiol*, *53*(5), 1685–92. doi:10.1128/JCM.00323-15