SeroTools Documentation

Release 0.2.1

Joseph D. Baugher, Ph.D.

Sep 04, 2020

Contents

Contents:

CHAPTER 1

SeroTools

This package serves as a toolkit and repository for the White-Kauffmann-Le Minor scheme for *Salmonella* serotyping, which is made available in multiple formats, along with methods for querying and comparing serovar names and antigenic formulae, as well as determining the most abundant serovar for a cluster of isolates.

SeroTools was developed by the United States Food and Drug Administration, Center for Food Safety and Applied Nutrition.

- Free software
- Documentation: https://serotools.readthedocs.io
- Source Code: https://github.com/CFSAN-Biostatistics/serotools
- PyPI Distribution: https://pypi.python.org/pypi/serotools

1.1 Introduction

Salmonella bacteria are major foodborne pathogens estimated by the U.S. Centers for Disease Control and Prevention to cause 1.35 million infections annually in the United States[1]. Serological subtyping (serotyping) of *Salmonella* has historically been a critical component of characterization and successful outbreak identification and traceback efforts employed by public health researchers and regulatory agencies. The White-Kauffmann-Le Minor (WKL) *Salmonella* serotyping scheme specifies the commonly accepted naming and formatting conventions for *Salmonella* serotyping data and the antigenic factors (and other characteristics) which define each serovar. *Salmonella* serotyping data is routinely employed by a broad range of scientific researchers, physicians, public health professionals, food safety experts, etc.

SeroTools addresses multiple critical needs for the efficient analysis of *Salmonella* serotyping data. As technological advances continue to produce a range of high resolution subtyping options, including *in silico* serovar prediction based on whole genome sequencing, new tools are necessary for efficient method-comparison studies and quality control applied to increasingly large numbers of isolates. SeroTools serves as the only multiformat WKL repository

[1] The U.S. Centers for Disease Control and Prevention. <https://www.cdc.gov/salmonella/index.html>.

accessible for software development and provides the only existing tools for querying the WKL scheme, comparing serovars for congruence, and predicting the most abundant serovar for clusters of isolates.

1.2 Features

- Query the White-Kauffmann-Le Minor *Salmonella* serotyping repository
- Compare serovar predictions for state of congruence
- Determine the most abundant serovar for a cluster of isolates

1.3 Citing SeroTools

To cite SeroTools, please reference the SeroTools GitHub repository:

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

1.4 License

See the LICENSE file included in the SeroTools distribution.

CHAPTER 2

Installation

At the command line:

```
$ pip install --user serotools
```

Or, if you have virtualenvwrapper installed:

```
$ mkvirtualenv serotools
$ pip install serotools
```

2.1 Upgrading SeroTools

If you previously installed with pip, you can upgrade to the newest version from the command line:

```
$ pip install --user --upgrade serotools
```

2.2 Uninstalling SeroTools

If you installed with pip, you can uninstall from the command line:

```
$ pip uninstall serotools
```

CHAPTER 3

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 which are being adopted by public health agencies for quality control and as an alternative to serological testing. 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 White-Kauffmann-Le Minor (WKL) repository accessible for software development. 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.

CHAPTER 4

Input Formatting

The input data must follow the nomenclature conventions as specified in[1].

- Named serovars (subsp. *enterica*) generally compose a single word or concatenation (e.g. Saintpaul) with no whitespace, with a few exceptions (e.g. Paratyphi B, II Alsterdorf), and should not contain hyphens or additional information (e.g. Choleraesuis is correct, while Cholerae-suis and Cholerae suis are incorrect).
- Named variants specified in[1] must adhere to the format below. For example, Westhampton var. 15+ or Westhampton var. 15+,34+:

```
<SerovarName> var. <factor>+,<factor>+
```

- The proper convention for an antigenic formula is as follows: a space must separate the subspecies symbol from the antigens; colons must separate the antigens; commas must separate the antigenic factors (e.g. I 1,4,[5],12:e,h:1,5:[R1...]):

```
<Subspecies> <O_factor1,O_factor2>:<P1_factor1,P1_factor2>:<P2_factor1,P2_factor2>
<Subspecies> <O_factor1,O_factor2>:<P1_factor1,P1_factor2>:<P2_factor1,P2_factor2>
↪:<otherH_factor1,otherH_factor2>
```

- Missing antigens should be specified using '–' (e.g. I 4,12,27:b:– or I 1,9,12:–:–).
- Optional, exclusive, and weakly agglutinable factors should be designated as follows:

```
optional            '[]'
exclusive           '{}'
weakly agglutinable '()'
```

- Input may contain the terms Nonmotile, Rough, or Mucoid, which will be converted into the appropriate format:

```
Nonmotile  :-:-
Rough      -:
Mucoid     -:
```

[1] Grimont PA, Weill FX. Antigenic Formulae of the Salmonella Serovars. 9th. Paris, France: WHO Collaborating Center for Reference and Research on Salmonella, Institut Pasteur; 2007 <https://www.pasteur.fr/sites/default/files/veng_0.pdf>.

- Phage conversion factors denoted by underlining in[1] are denoted as optional '[]' in SeroTools, with the exception of exclusive phage conversion factors (e.g. {15} and {15,34}).

CHAPTER 5

Usage

SeroTools provides methods for querying and comparing serovar names and antigenic formulae, as well as determining the most abundant serovar for a cluster of isolates.

5.1 query

Query the White-Kauffmann-Le Minor (WKL) repository by submitting one of more serovar names or antigenic formulas in an input file composed of a single query per line:

```
$ serotools query -i <input_file>
```

or as a command line argument:

```
$ serotools query -s 'Paratyphi A'
```

Output:

```
Input        Name         Formula             Match
Paratyphi A  Paratyphi A  I [1],2,12:a:[1,5]  exact
```

5.2 compare

Compare serovar predictions by evaluating multiple states of congruence (exact, congruent, minimally congruent, incongruent). Serovar names and/or antigenic formulae may be submitted in a tab-delimited input file composed of two columns of serovar predictions:

```
$ serotools compare -i <input_file>
```

or as command line arguments:

```
$ serotools compare -1 'Hull' -2 'I 16:b:1,2'
```

Output:

```
Serovar1    Name    Formula     Serovar2    Name    Formula     Result
Hull        Hull    I 16:b:1,2  I 16:b:1,2  Hull    I 16:b:1,2  exact
```

5.3 cluster

Determine the most abundant serovar(s) for one or more clusters of isolates. Input data must be submitted in the form of a tab-delimited file in which each line consists of a cluster ID and one serovar as follows:

Input File - example.txt:

```
cluster1    Dunkwa
cluster1    Dunkwa
cluster1    Utah
cluster2    Hull
```

```
$ serotools cluster -i example.txt
```

Output:

```
ClusterID   ClusterSize Input   Name    Formula      P_Exact  P_Congruent   P_MinCon
cluster1    2           Dunkwa  Dunkwa  I 6,8:d:1,7  0.6667   0.6667        0.6667
cluster2    1           Hull    Hull    I 16:b:1,2   1.0      1.0           1.0
```

CHAPTER 6

Congruence

SeroTools evaluates multiple levels of congruence for comparisons between serovar designations.

6.1 exact

Exact matches must meet one of the following criteria:

- Two serovar designations are the identical string:

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

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

```
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
```

- Neither serovar designation includes any antigenic factors, and the subspecies designations match:

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

6.2 congruent

Congruent matches must meet the following criteria:

- The subspecies field must be present for both serovars or neither.
- All **required** antigenic factors match. 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:-

6.3 minimally congruent

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). 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

- Note - the empty set (–) is a subset of every set

The minimally congruent designation is unique to SeroTools and is useful for distinguishing between two scenarios:

- Serovars which differ due to sample misannotation (incongruent)
- Serovars derived from correctly annotated samples with variation based solely on missing information. When comparing serovar designations, minor differences may be expected due to method-specific irregularities, for example reagent variation for laboratory-based techniques or the presence of nonproductive genomic data when comparing antigenic agglutination to *in silico*-based 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.

6.4 incongruent

Any comparison which is not 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

CHAPTER 7

Repository

SeroTools provides a repository of the White-Kauffmann-Le Minor (WKL) Salmonella serotyping scheme based on these *References* in the following formats:

- Python data structures (serotools.py)
  - pandas DataFrame:

```
wklm_df
```

  - Dictionaries:

```
wklm_name_to_formula
wklm_formula_to_name
```

  - Lists with common indexing:

```
wklm_name
std_wklm_name (standardized for matching)
wklm_formula
std_wklm_formula (standardized for matching)
wklm_sp (species)
wklm_subsp (subspecies)
wklm_O
wklm_P1
wklm_P2
wklm_other_H
wklm_group (O group)
wklm_old_group (previous O group)
```

- An Excel spreadsheet (White-Kauffman-LeMinor-Scheme.xlsx)
- A tab-delimited text file (White-Kauffman-LeMinor_scheme.tsv)

CHAPTER 8

Notes

1. Phage conversion factors denoted with underlining in[1] are here denoted as optional '[]' with the exception of the exclusive factors (e.g. {15} and {15,34}).
2. Serovar Montevideo is listed twice in[1]: O:7 I 6,7,[14]:g,m,[p],s:[1,2,7] and O:54 I {6,7,[14]}{54}:g,m,s:–. The profile from the 'Alphabetical List' p. 137 will be used here - I 6,7,[14],[54]:g,m,[p],s:[1,2,7].
3. As in[1], although S. bongori is not a subspecies of S. enterica, symbol 'V' was retained in order to avoid formatting confusion.

[1] Grimont PA, Weill FX. Antigenic Formulae of the Salmonella Serovars. 9th. Paris, France: WHO Collaborating Center for Reference and Research on Salmonella, Institut Pasteur; 2007 <https://www.pasteur.fr/sites/default/files/veng_0.pdf>.

CHAPTER 9

References

SeroTools includes serovar names and antigenic formulas as specified in the following publications:

1. Grimont PA, Weill FX. Antigenic Formulae of the Salmonella Serovars. 9th. Paris, France: WHO Collaborating Center for Reference and Research on Salmonella, Institut Pasteur; 2007 <https://www.pasteur.fr/sites/default/files/veng_0.pdf>.
2. Guibourdenche M, Roggentin P, Mikoleit M, Fields PI, Bockemühl J, Grimont PA, Weill FX. Supplement 2003-2007 (No. 47) to the White-Kauffmann-Le Minor scheme. Res Microbiol. 2010 Jan-Feb;161(1):26-9 <https://doi.org/10.1016/j.resmic.2009.10.002>.
3. Issenhuth-Jeanjean S, Roggentin P, Mikoleit M, Guibourdenche M, de Pinna E, Nair S, Fields PI, Weill FX. Supplement 2008-2010 (no. 48) to the White-Kauffmann-Le Minor scheme. Res Microbiol. 2014 Sep;165(7):526-30 <https://doi.org/10.1016/j.resmic.2014.07.004>.
4. Bugarel M, den Bakker HC, Nightingale KK, Brichta-Harhay DM, Edrington TS, Loneragan GH. Two Draft Genome Sequences of a New Serovar of Salmonella enterica, Serovar Lubbock. Genome Announc. 2015 Apr 16;3(2) <https://doi.org/10.1128/genomeA.00215-15>.

CHAPTER 10

Credits

10.1 Author

- Joseph D. Baugher, Ph.D. <joseph.baugher@fda.hhs.gov>

10.2 Development Lead

- Joseph D. Baugher, Ph.D. <joseph.baugher@fda.hhs.gov>

10.3 CFSAN Bioinformatics Team

- Joseph D. Baugher, Ph.D. <joseph.baugher@fda.hhs.gov>

10.4 External Contributors

CHAPTER 11

History

11.1 0.2.1 (2020-09-04)

- Updated documentation
- Added JOSS manuscript

11.2 0.2.0 (2020-02-17)

Significant updates in this version - not backwards compatible.

- The underlying data structures have been converted to pandas Series and DataFrames.
- New 'cluster' subcommand functionality provides the most abundant serovar(s) for clusters of isolates.
- The 'predict' subcommand functionality has been merged into the 'query' subcommand, such that the default query will return any exact, congruent, and minimally congruent matches unless only exact matches are desired.
- The WKL repository is now available as a pandas DataFrame, in addition to dictionaries and lists.

11.3 0.1.1 (2019-11-27)

- Corrected a variable name in cli.py
- Updated the algorithm for minimally congruent serovars

11.4 0.1.0 (2019-11-19)

- Initial version.

CHAPTER 12

Indices and tables

- genindex
- search