1 Serotype detection using SeroBA

1.1 Introduction

SeroBA is a software tool for identifying the serotype of samples from Illumina reads. This tutorial will walk you through using SeroBA for serotyping of *Streptococcus pneumoniae* samples.

For more in depht information about SeroBA, please refer to the paper:

> **SeroBA: rapid high-throughput serotyping of Streptococcus pneumoniae from whole genome sequence data** Epping L, van Tonder, AJ, Gladstone RA, GPS Consortium, Bentley SD, Page AJ, Keane JA *bioRxiv preprint, 2017 Sep.; doi: 10.1101/179465*

1.2 Learning outcomes

By the end of this tutorial you can expect to be able to:

- Understand serotyping, why it is important and what it can be used for
- Run SeroBA on several samples to predict their serotype
- Summarise the SeroBA results for several samples
- Interpret the detailed output of SeroBA
- Download and prepare the S. pneumoniae databases from PneumoCAT for use with SeroBA

1.3 Tutorial sections

This tutorial comprises the following sections:

1. What is serotyping?
2. Preparation of databases before running SeroBA
3. Running SeroBA
4. Interpreting the results of SeroBA

1.4 Authors

This tutorial was created by Sara Sjunnebo.

1.5 Running the commands from this tutorial

You can run the commands in this tutorial either directly from the Jupyter notebook (if using Jupyter), or by typing the commands in your terminal window.

1.5.1 Running commands on Jupyter

If you are using Jupyter, command cells (like the one below) can be run by selecting the cell and clicking *Cell -> Run* from the menu above or using *ctrl Enter* to run the command. Let's give this a try by printing our working directory using the *pwd* command and listing the files within it. Run the commands in the two cells below.

```
pwd
```

```
ls -l
```

1.5.2 Running commands in the terminal

You can also follow this tutorial by typing all the commands you see into a terminal window. This is similar to the "Command Prompt" window on MS Windows systems, which allows the user to type DOS commands to manage files.

To get started, select the cell below with the mouse and then either press control and enter or choose Cell -> Run in the menu at the top of the page.

```
echo cd $PWD
```

Now open a new terminal on your computer and type the command that was output by the previous cell followed by the enter key. The command will look similar to this:

```
cd /home/manager/pathogen-informatics-training/Notebooks/SEROBA/
```

Now you can follow the instructions in the tutorial from here.

1.6 Let's get started!

This tutorial assumes that you have SeroBA installed on your computer. For download and installation instructions, please see the SeroBA GitHub-page.

To check that you have installed the software correctly, you can run the following command:

```
seroba -help
```

This should return the following help message:

```
usage: seroba <command> <options>

optional arguments:
  -h, --help     show this help message and exit

Available commands:

    getPneumocat
                 downloads genetic information from PneumoCat
    createDBs    creates Databases for kmc and ariba
    runSerotyping
                 indetify serotype of your input data
    summary      output folder has to contain all folders with prediction
                 results
    version      Get versions and exit
```

To get started with the tutorial, head to the first section: What is serotyping? The answers to all questions in the tutorial can be found here.

2 Serotyping

A species can be subdivided into different groups based on the antigens expressed on their cell surface. These groups are called serotypes or serovars and the different properties between them can vary greatly. For example, which antigens are expressed on the cell surface of a bacterium can make it more or less virulent, or more or less sensitive to substances like antibiotics.

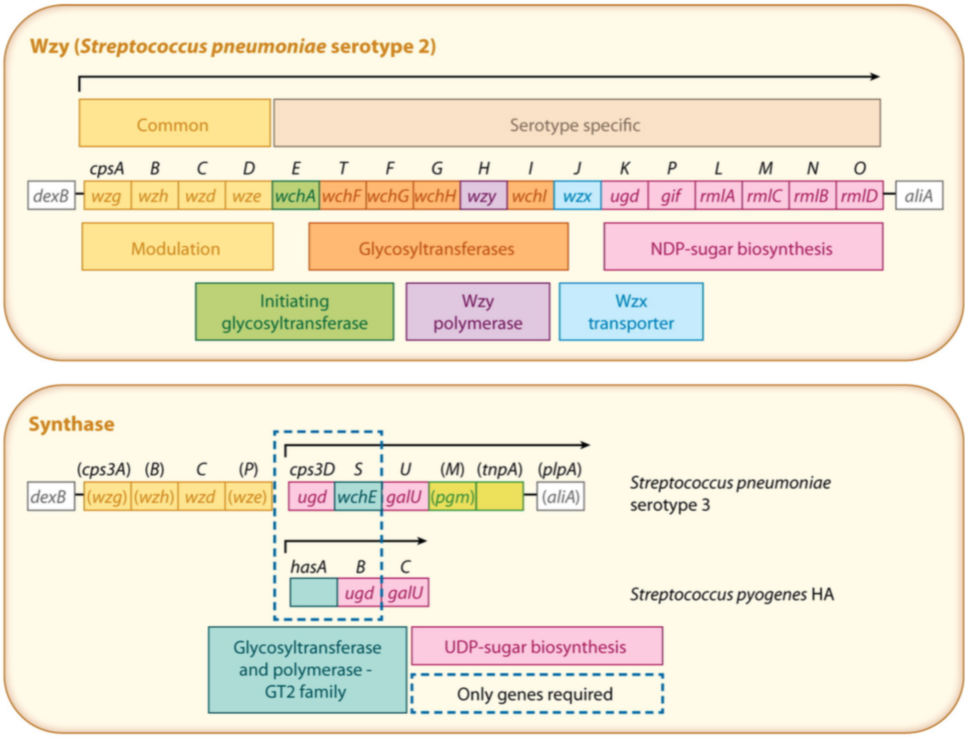
2.1 *Streptococcus pneumoniae*

Diseases that are caused by *Streptococcus pneumoniae* are a big problem in public health across the world. There are around 100 known serotypes of *S. pneumoniae*. The current conjugate vaccine (pcv13) covers the 13 most common serotypes causing invasive pneumococcal infections in industrialised countries, but because vaccines are serotype specific, it is of great value to be able to quickly and accurately determine serotypes in order to monitor epidemiological trends of *S. pneumoniae* following the introduction of effective vaccines.

2.2 Serotyping *S. pneumoniae*

The serotype of a strain of *S. pneumoniae* is determined by the capsular polysaccharide biosynthesis (cps) locus, pictured below. It is a major virulence factor in *S. pneumoniae*, encoding polysaccharide chains that form a capsule around the cell, helping the bacterium avoid the human immune system.

The cps locus can be very similar between serotypes and based on this serotypes can be grouped into serogroups. Serotypes can also be grouped into serogroups by how similar the anigenic response they trigger is.



Wzy (Streptococcus pneumoniae serotype 2)
Common
Serotype specific
cpsA B C D E T F G H I J K P L M N O
dexB wzg wzh wzd wze wchA wchF wchG wchH wzy wchI wzx ugd gif rmlA rmlC rmlB rmlD aliA
Modulation
Glycosyltransferases
NDP-sugar biosynthesis
Initiating glycosyltransferase
Wzy polymerase
Wzx transporter
Synthase
(cps3A) (B) C (P) cps3D S U (M) (tnpA) (plpA)
dexB (wzg) (wzh) wzd (wze) ugd wchE galU (pgm) (aliA)
Streptococcus pneumoniae serotype 3
hasA B C
ugd galU
Streptococcus pyogenes HA
Glycosyltransferase and polymerase - GT2 family
UDP-sugar biosynthesis
Only genes required

AR Yother J. 2011.
Annu. Rev. Microbiol. 65:563–81

cps locus

Traditionally, determining the serotype of *S. pneumoniae* has predominantly been done with the Quellung reaction or PCR, each with their own limitations. Lately, focus has shifted towards serotyping directly from genomic data.

2.3 SeroBA

Existing software to infer serotypes from genomic data are limited and do not scale well. SeroBA is a pipeline that can quickly and accurately determine the serotype of *S. pneumoniae* from WGS data (Illumina paired-end reads). It uses k-mer analysis and references to determine the serotype of a sample.

In this tutorial we will walk you through how to determine the serotypes of two samples using SeroBA, from setting up the necessary databases, to running the analysis, and finally how to interpret the results.

For more information and to explore the code behind SeroBA, you can visit the GitHub page.

We will start with setting up the databases with a few simple commands. You can also return to the index.

3 Preparation of databases before running SeroBA

In order to use SeroBA for serotyping we must first download and prepare the necessary databases. Start by moving into the data directory:

```
cd data
```

Now download the database from the GitHub repository:

```
svn checkout "https://github.com/sanger-pathog\
           ens/seroba/trunk/database"
```

NOTE if you are running a version of SeroBA older than v.0.1.3 the database is not packaged with the program and you will have to download it using the below command instead:

```
seroba getPneumocat database_dir
```

KMC is used by SeroBA to count k-mers and ARIBA is used to avoid the need for reads to be mapped to all reference sequences. Both of these require a database to be set up.

To create a database for KMC and ARIBA run **createDBs**:

```
seroba createDBs database/ kmer_size
```

Where the options are:

```
database       The database directory which you just downloaded
kmer_size      The k-mer size you want to use for kmc. Recommended = 71
```

SeroBA uses a default k-mer size of 71 for a read length of 250 bp. When deciding on a k-mer size, it is worth knowing that while a smaller k-mer size can keep the memory requirements low, it will also reduce the specificity. On the other hand, a larger k-mer size will require a larger amount of memory but will produce more unique k-mers and thus increase the specificity. What k-mer size to use also depends on the read length.

```
seroba createDBs database/ 71
```

If you are working with SeroBA on the Sanger farm, the database with k-mer size 71 is already available centrally. This means you do not need to create the database for using SeroBA on the Sanger farm.

However, for the sake of this tutorial, the above steps need to be compleated before you can continue with the tutorial.

In the next section we are going to run SeroBA to determine the serotype of one sample. You can also return to the index or revisit the previous section.

4 Running SeroBA

We are now ready to use SeroBA to determine the serotype of our samples. Move into the data directory where we keep the reads. In this case we have called the directory **run_seroba** and it is in the directory called **data**.

```
cd data/run_seroba
```

Have a quick look at the contents of the directory:

```
ls -al
```

As you can see, there are two gzipped fastq files for each sample, one for forward reads and one for reverse reads.

Now run SeroBA using the **runSerotyping** command:

`seroba runSerotyping database forward_read reverse_read prefix`

Where the options are:

database	path to the database directory
forward_read	forward read file in fastq format
reverse_read	reverse read file in fastq format
prefix	a unique prefix

```
seroba runSerotyping ../database/ sample1_1.fq.gz \
           sample1_2.fq.gz sample1
```

If you are running Seroba on the Sanger farm and instead want to use the central database, you can use:

```
seroba runSerotyping $SEROBA_DB forward_read reverse_read prefix
```

Lets have a look at the results in the next section. You can also return to the index or revisit the previous section.

5 Interpreting the results

Now lets have a look at the results. Move into the data directory again.

```
cd data/run_seroba
```

Look at the file called pred.tsv in your results directoy, sample1.

```
cat sample1/pred.tsv
```

You can see three columns. The first one contains the prefix you chose for the run. The second one contains the predicted serotype and the third column may contain a comment regarding contamination. So, in this case we can see that sample1 was predicted to be of serotype 8 and at least 10% of the reads are called as a different snp than the other reads i.e. there is contamination.

Now, let's try again with the rest of the samples. All we need to do is to run the runSerotyping option, but this time we will do it in a for-loop so that we do not have to run the command manually for each sample.

The command we are going to use will run seroba runSerotyping on all fastq files in the working directory, so to avoid serytyping sample1 again, first move the fastq files for this sample out from the working directory:

```
mv sample1_* ../
```

Run the command below. It will take around 10 minutes.

```
for file in *_1.fq.gz; do seroba runSerotyping \
           ../database/ $file "${file//_1.fq.gz/_2.fq.gz}" \
           "${file//_1.fq.gz/}"; done
```

Now that we have performed multiple runs, we might want to create a summary of the results. To do this, move up one level to the data directory.

```
cd ..
```

Now run the **summary** option:

`seroba summary results_dir`

Where the results_dir in this case is run_seroba.

```
seroba summary run_seroba
```

Have a look at the resulting tsv file.

```
cat summary.tsv
```

As you can see, you now have a summary of all the runs in one file. One sample that might require some explanation is sample3 which was serotyped as 6E(6B). Serogroup 6 is a bit different from the other serogroups. The serotype of a sample can be 6E on the genotypic level, and 6A or 6B on the phenotypic level. In this case, sample3 has the 6E genotype and the 6B phenotype.

Now follows some questions for you to go over what you have learned in this tutorial one more time! You can find the answers in the link at the bottom of the page.

6 Questions

1. Your data requires you to use a kmer size of 60. You have already downloaded the database to a directory called database_60/. What command would you run to create a database for KMC and ARIBA with kmer size 60?

2. What is the predicted serotype of sample7?

3. How would you interpret the comment for sample7?

This is the end of the tutorial. You can return to the index or revisit the previous section.

The answers to the questions in this tutoial can be found here.