Swine Flu Epitope Analysis

Tan Swan

4/29/2018

Project Background

There is a study on swine vaccination and the project collaborators need to produce an inactivated vaccine to immunize pigs. Prior to do so, the pigs have to be challenged with a virus strain in order to test the efficacy of experimental inactivated vaccine.

Taking the previous published swine DNA vaccine as reference, the collaborator intended to know which circulating swine influenza A virus (IAV) is suitable to serve as the challenge strain.

Post-Analysis Goals

1. To screen H1N1 IAV sequences (provided by the collaborator) against the epitopes in the DNA vaccine to find good matches using two approaches: Epitope Content Comparison (EpiCC) and JanusMatrix (JMX).
2. Rank the viruses to select top matches to be candidate for the challenge strain.

Pre-checkpoint - Libraries to include

```
library(tidyverse)
library(gridExtra)
library(ggplot2)
library(plotly)
library(knitr)
```

Part I

About EpiCC

Epitope Content Comparison (EpiCC) - A web-based computational method that facilitates pairwise comparison of protein sequences based on immunological property, i.e. T cell epitope content, rather than sequence identity, and evaluated its ability to classify swine influenza A virus (IAV) strain relatedness to estimate cross-protective potential of a vaccine strain for circulating viruses (Gutiérrez et. al, 2017).

Aim

To identify strains that have highest epitope content relatedness to the reference vaccine strain of MHC Class I/II (VACCINE_EPITOPES_CLASSI-NTC8684-ERNA41H-7SOJI/VACCINE_EPITOPES_CLASSII-NTC8682-ERNA41H-1SOJII) across H1N1 Swine IAV whole genome, i.e. a total of 8 protein segments/antigens.

Materials

Output from EpiCC tool to process and analyze: 8 outfiles for MHC Class I and II respectively.

Methodology/ Working steps

1. Load EpiCC MHC Class I data and clean up unnecessary rows and columns

Quick view of the processed data from one of the protein segments

```
head(colnames(epicc_data_1_clsI))
```

```
## [1] "id"
## [2] "GB_KY888027-A_SWINE_KANSAS_A01378019_2017-SEGMENT_1"
## [3] "GB_KY970162-A_SWINE_MICHIGAN_A01259076_2017-SEGMENT_1"
## [4] "GB_MF116355-A_SWINE_KANSAS_A01378027_2017-SEGMENT_1"
## [5] "GB_MF373215-A_SWINE_IOWA_A01672518_2017-SEGMENT_1"
## [6] "GB_MF373233-A_SWINE_NEBRASKA_A01672345_2017-SEGMENT_1"
```

```
nrow(epicc_data_1_clsI)
```

```
## [1] 1
```

```
ncol(epicc_data_1_clsI)
```

```
## [1] 74
```

2. Transform data from wide form to long form, perform sorting and extracting top 10 EpiCC score from each protein

Showing a working example from one of the protein segments: PB2

```
epicc_data_1_clsI_long <- epicc_data_1_clsI %>%
  gather(key = strains, value = PB2_score,
         `GB_KY888027-A_SWINE_KANSAS_A01378019_2017-SEGMENT_1`:`VACCINE_EPITOPES_CLASSI-NTC8684-ERNA41H-7SOJI`) %
>%
  select(strains,PB2_score) %>% mutate(header = strains) %>%
  separate(header, into = c("ID", "Sequence", "Segment"), sep = "-", extra = "merge")
epicc_data_1_clsI_long$PB2_score <- as.numeric(epicc_data_1_clsI_long$PB2_score)
epicc_data_1_clsI_sort <- epicc_data_1_clsI_long %>% filter(!Sequence == "NTC8684") %>%
  arrange(Sequence) %>% arrange(desc(PB2_score))
epicc_data_1_clsI_top10 <- head(epicc_data_1_clsI_sort, 10)
epicc_data_1_clsI_top10 <- epicc_data_1_clsI_top10 %>% select(Sequence, Segment)
epicc_data_1_clsI_top10
```

```
##                                Sequence   Segment
## 1           A_SWINE_TEXAS_A02214607_2017 SEGMENT_1
## 2            A_SWINE_IOWA_A01667091_2017 SEGMENT_1
## 3        A_SWINE_OKLAHOMA_A02214419_2017 SEGMENT_1
## 4  A_SWINE_NORTH_CAROLINA_A01672751_2017 SEGMENT_1
## 5            A_SWINE_IOWA_A01104104_2017 SEGMENT_1
## 6        A_SWINE_NEBRASKA_A02216645_2017 SEGMENT_1
## 7          A_SWINE_KANSAS_A01378019_2017 SEGMENT_1
## 8          A_SWINE_KANSAS_A01378038_2017 SEGMENT_1
## 9            A_SWINE_IOWA_A01672518_2017 SEGMENT_1
## 10 A_SWINE_NORTH_CAROLINA_A01672011_2017 SEGMENT_1
```

2a. Exploratory analysis of H1N1 Swine IAV dataset

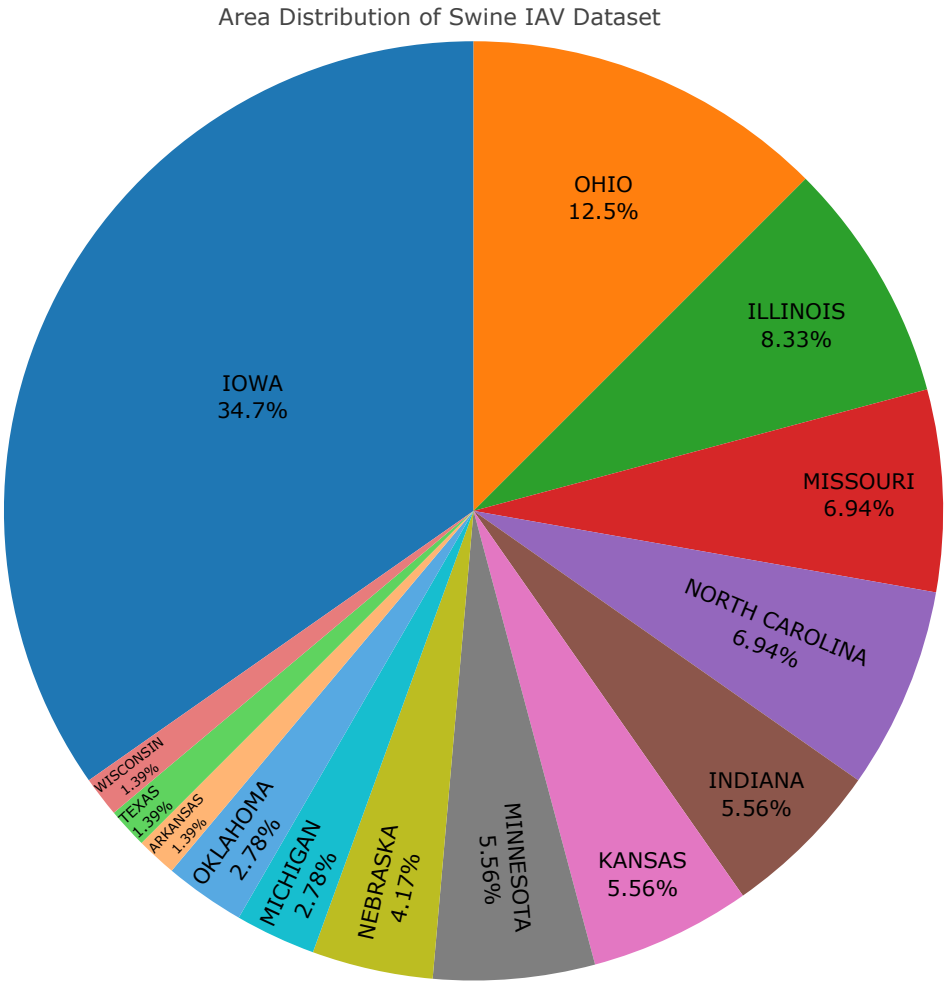

Figure 1.1 | The pie chart shows area distribution of swine IAV dataset. More than 30% of swine IAV strains are from Iowa. Second abundance being OHIO, followed by Illinois, Missouri and North Carolina.

3. Combine top 10 EpiCC score of each protein

MHC Class I top 10 data preview

```
head(epicc_data_clsI_top10_all)
```

```
##                               Sequence   Segment
## 1          A_SWINE_TEXAS_A02214607_2017 SEGMENT_1
## 2           A_SWINE_IOWA_A01667091_2017 SEGMENT_1
## 3       A_SWINE_OKLAHOMA_A02214419_2017 SEGMENT_1
## 4 A_SWINE_NORTH_CAROLINA_A01672751_2017 SEGMENT_1
## 5           A_SWINE_IOWA_A01104104_2017 SEGMENT_1
## 6       A_SWINE_NEBRASKA_A02216645_2017 SEGMENT_1
```

```
tail(epicc_data_clsI_top10_all)
```

```
##                            Sequence   Segment
## 75 A_SWINE_MINNESOTA_A02214666_2017 SEGMENT_8
## 76    A_SWINE_KANSAS_A01378027_2017 SEGMENT_8
## 77      A_SWINE_IOWA_A02215041_2017 SEGMENT_8
## 78      A_SWINE_IOWA_A02216456_2017 SEGMENT_8
## 79      A_SWINE_OHIO_A02219547_2017 SEGMENT_8
## 80   A_SWINE_INDIANA_A02216644_2017 SEGMENT_8
```

```
unique(epicc_data_clsI_top10_all$Sequence)
```

```
##  [1] "A_SWINE_TEXAS_A02214607_2017"
##  [2] "A_SWINE_IOWA_A01667091_2017"
##  [3] "A_SWINE_OKLAHOMA_A02214419_2017"
##  [4] "A_SWINE_NORTH_CAROLINA_A01672751_2017"
##  [5] "A_SWINE_IOWA_A01104104_2017"
##  [6] "A_SWINE_NEBRASKA_A02216645_2017"
##  [7] "A_SWINE_KANSAS_A01378019_2017"
##  [8] "A_SWINE_KANSAS_A01378038_2017"
##  [9] "A_SWINE_IOWA_A01672518_2017"
## [10] "A_SWINE_NORTH_CAROLINA_A01672011_2017"
## [11] "A_SWINE_NEBRASKA_A02219793_2017"
## [12] "A_SWINE_INDIANA_A01672825_2017"
## [13] "A_SWINE_ILLINOIS_A02218178_2017"
## [14] "A_SWINE_IOWA_A02216046_2017"
## [15] "A_SWINE_ILLINOIS_A01932036_2017"
## [16] "A_SWINE_ILLINOIS_A01672343_2017"
## [17] "A_SWINE_ILLINOIS_A02214663_2017"
## [18] "A_SWINE_IOWA_A02214835_2017"
## [19] "A_SWINE_MINNESOTA_A02214666_2017"
## [20] "A_SWINE_IOWA_A02215202_2017"
## [21] "A_SWINE_KANSAS_A01378027_2017"
## [22] "A_SWINE_IOWA_A01667089_2017"
## [23] "A_SWINE_IOWA_A02221505_2017"
## [24] "A_SWINE_IOWA_A02217282_2017"
## [25] "A_SWINE_IOWA_A02215038_2017"
## [26] "A_SWINE_ILLINOIS_A02219783_2017"
## [27] "A_SWINE_OHIO_17TOSU1384_2017"
## [28] "A_SWINE_OHIO_17TOSU1386_2017"
## [29] "A_SWINE_OHIO_A01354304_2017"
## [30] "A_SWINE_OHIO_A01354305_2017"
## [31] "A_SWINE_OKLAHOMA_A01672680_2017"
## [32] "A_SWINE_MISSOURI_A01932424_2017"
## [33] "A_SWINE_NORTH_CAROLINA_A01785281_2017"
## [34] "A_SWINE_KANSAS_A01378037_2017"
## [35] "A_SWINE_IOWA_A02215041_2017"
## [36] "A_SWINE_IOWA_A02216456_2017"
## [37] "A_SWINE_OHIO_A02219547_2017"
## [38] "A_SWINE_INDIANA_A02216644_2017"
```

```
unique(epicc_data_clsI_top10_all$Segment)
```

```
## [1] "SEGMENT_1" "SEGMENT_2" "SEGMENT_3" "SEGMENT_4" "SEGMENT_5" "SEGMENT_6"
## [7] "SEGMENT_7" "SEGMENT_8"
```

The numbering in each segment stands for type of protein encoded in flu genome. 1:PB2, 2: PB1, 3:PA, 4:HA, 5:NP, 6:NA, 7:M, 8:NS

4. Identify top occuring strains

```
most_occurence_strain_clsI <- epicc_data_clsI_top10_all %>% group_by(Sequence) %>% summarise(count = n()) %>% arr
ange(desc(count)) %>% filter(count >= 3)
```

```
max(most_occurence_strain_clsI$count)
```

```
## [1] 5
```

```
most_occurence_strain_clsI
```

```
## # A tibble: 12 x 2
##                          Sequence count
##                             <chr> <int>
##  1     A_SWINE_IOWA_A01672518_2017     5
##  2     A_SWINE_IOWA_A02215202_2017     5
##  3     A_SWINE_IOWA_A02221505_2017     5
##  4     A_SWINE_IOWA_A01104104_2017     4
##  5     A_SWINE_IOWA_A01667091_2017     4
##  6     A_SWINE_IOWA_A02214835_2017     4
##  7     A_SWINE_IOWA_A02215038_2017     4
##  8   A_SWINE_KANSAS_A01378027_2017     4
##  9    A_SWINE_TEXAS_A02214607_2017     4
## 10 A_SWINE_ILLINOIS_A01932036_2017     3
## 11   A_SWINE_KANSAS_A01378038_2017     3
## 12 A_SWINE_NEBRASKA_A02216645_2017     3
```

Maximum frequency = 5 out of 8 (if the EpiCC score of a strain remain in the top 10 list of all proteins, then it will have frequency of 8). In this case, since the maximum frequency is only 5, this means that these strains are only found in the top 10 list of 5 proteins. Here, we considered strains that of frequency between 3 - 5.

5. Repeat steps 1 - 4 for MHC Class II data

MHC Class II top 10 data preview

```
head(epicc_data_clsII_top10_all)
```

```
##                          Sequence   Segment
## 1   A_SWINE_KANSAS_A01378019_2017 SEGMENT_1
## 2 A_SWINE_MISSOURI_A01932424_2017 SEGMENT_1
## 3   A_SWINE_KANSAS_A01378027_2017 SEGMENT_1
## 4     A_SWINE_OHIO_A02219547_2017 SEGMENT_1
## 5 A_SWINE_NEBRASKA_A02219793_2017 SEGMENT_1
## 6 A_SWINE_OKLAHOMA_A02214419_2017 SEGMENT_1
```

```
tail(epicc_data_clsII_top10_all)
```

```
##                                Sequence   Segment
## 75         A_SWINE_KANSAS_A01378027_2017 SEGMENT_8
## 76       A_SWINE_NEBRASKA_A02219793_2017 SEGMENT_8
## 77 A_SWINE_NORTH_CAROLINA_A01785281_2017 SEGMENT_8
## 78         A_SWINE_KANSAS_A01378019_2017 SEGMENT_8
## 79       A_SWINE_MISSOURI_A02216048_2017 SEGMENT_8
## 80           A_SWINE_IOWA_A02221506_2017 SEGMENT_8
```

```
unique(epicc_data_clsII_top10_all$Sequence)
```

```
##  [1] "A_SWINE_KANSAS_A01378019_2017"
##  [2] "A_SWINE_MISSOURI_A01932424_2017"
##  [3] "A_SWINE_KANSAS_A01378027_2017"
##  [4] "A_SWINE_OHIO_A02219547_2017"
##  [5] "A_SWINE_NEBRASKA_A02219793_2017"
##  [6] "A_SWINE_OKLAHOMA_A02214419_2017"
##  [7] "A_SWINE_IOWA_A01672342_2017"
##  [8] "A_SWINE_KANSAS_A01378038_2017"
##  [9] "A_SWINE_ILLINOIS_A02218178_2017"
## [10] "A_SWINE_IOWA_A01667088_2017"
## [11] "A_SWINE_NORTH_CAROLINA_A01672751_2017"
## [12] "A_SWINE_NORTH_CAROLINA_A01785281_2017"
## [13] "A_SWINE_IOWA_A01672824_2017"
## [14] "A_SWINE_IOWA_A02214479_2017"
## [15] "A_SWINE_MISSOURI_A02214279_2017"
## [16] "A_SWINE_NORTH_CAROLINA_A01785282_2017"
## [17] "A_SWINE_ILLINOIS_A01932036_2017"
## [18] "A_SWINE_MISSOURI_A02216048_2017"
## [19] "A_SWINE_IOWA_A02221508_2017"
## [20] "A_SWINE_OKLAHOMA_A01672680_2017"
## [21] "A_SWINE_NORTH_CAROLINA_A01672011_2017"
## [22] "A_SWINE_IOWA_A02215038_2017"
## [23] "A_SWINE_ARKANSAS_A02218161_2017"
## [24] "A_SWINE_TEXAS_A02214607_2017"
## [25] "A_SWINE_IOWA_A02215202_2017"
## [26] "A_SWINE_IOWA_A01104104_2017"
## [27] "A_SWINE_IOWA_A02214835_2017"
## [28] "A_SWINE_NEBRASKA_A02216645_2017"
## [29] "A_SWINE_IOWA_A01672518_2017"
## [30] "A_SWINE_IOWA_A02221505_2017"
## [31] "A_SWINE_IOWA_A02217282_2017"
## [32] "A_SWINE_KANSAS_A01378037_2017"
## [33] "A_SWINE_MINNESOTA_A02214666_2017"
## [34] "A_SWINE_NEBRASKA_A01672345_2017"
## [35] "A_SWINE_ILLINOIS_A02219783_2017"
## [36] "A_SWINE_IOWA_A02221506_2017"
```

```
unique(epicc_data_clsII_top10_all$Segment)
```

```
## [1] "SEGMENT_1" "SEGMENT_2" "SEGMENT_3" "SEGMENT_4" "SEGMENT_5" "SEGMENT_6"
## [7] "SEGMENT_7" "SEGMENT_8"
```

```
most_occurence_strain_clsII <- epicc_data_clsII_top10_all %>% group_by(Sequence) %>% summarise(count = n()) %>%
  arrange(desc(count)) %>% filter(count >= 3)
max(most_occurence_strain_clsII$count)
```

```
## [1] 7
```

```
most_occurence_strain_clsII
```

```
## # A tibble: 14 x 2
##                                Sequence count
##                                   <chr> <int>
##  1         A_SWINE_KANSAS_A01378027_2017     7
##  2       A_SWINE_ILLINOIS_A01932036_2017     6
##  3           A_SWINE_IOWA_A02215038_2017     4
##  4         A_SWINE_KANSAS_A01378038_2017     4
##  5 A_SWINE_NORTH_CAROLINA_A01785281_2017     4
##  6       A_SWINE_OKLAHOMA_A02214419_2017     4
##  7           A_SWINE_IOWA_A01104104_2017     3
##  8           A_SWINE_IOWA_A01672518_2017     3
##  9           A_SWINE_IOWA_A02215202_2017     3
## 10         A_SWINE_KANSAS_A01378019_2017     3
## 11         A_SWINE_KANSAS_A01378037_2017     3
## 12       A_SWINE_MISSOURI_A02216048_2017     3
## 13       A_SWINE_NEBRASKA_A02216645_2017     3
## 14       A_SWINE_OKLAHOMA_A01672680_2017     3
```

Maximum frequency = 7 out of 8. Same selection criteria as Class I data, we considered strains that of frequency starting 3.

6. To plot the frequency of MHC Class I/II top 10 strains from all 8 proteins

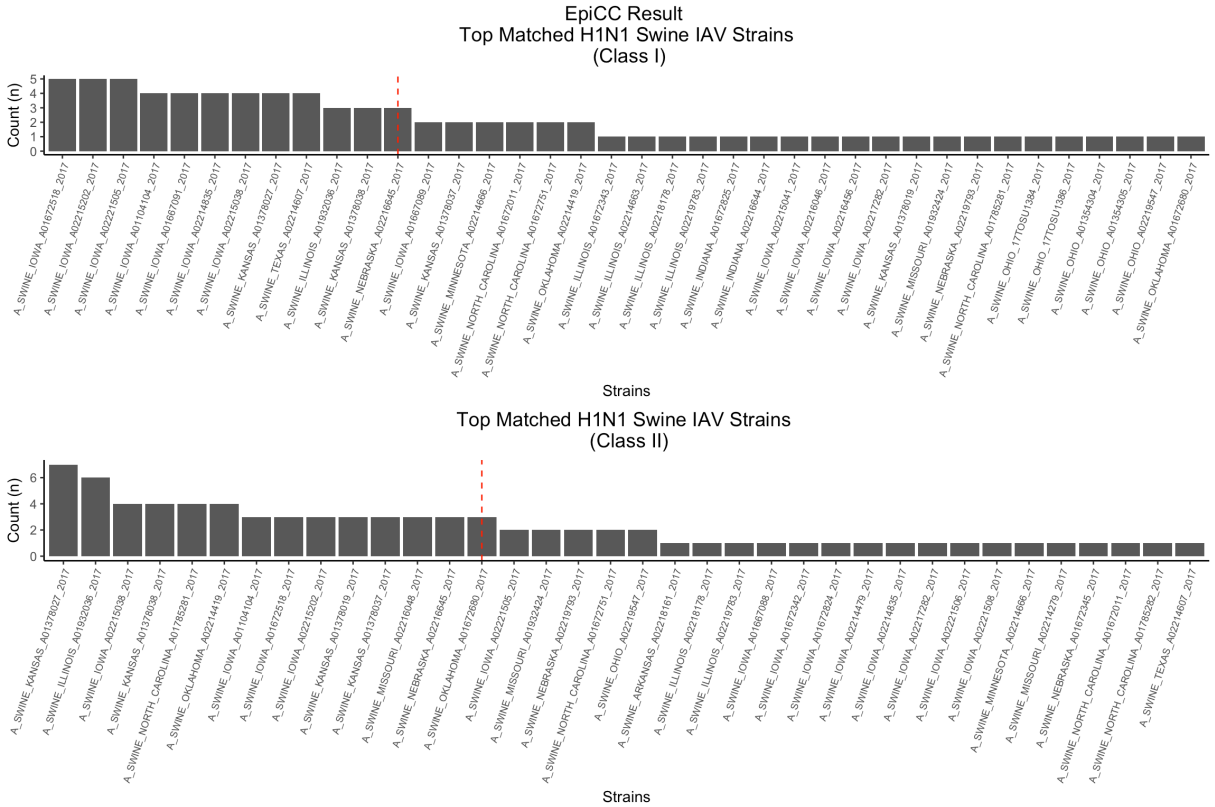

Figure 1.2 | Figure shows strains that are found in the top 10 list of every proteins and their frequencies were counted. This is to identify strains that are constantly having top EpiCC score across the whole genome. A reference line is drawn and strains are shortlisted based on the cut off point.

7. Identify strains that are common in both Class I and Class II

```
epicc_overlap <- most_occurence_strain_clsII$Sequence %in% most_occurence_strain_clsI$Sequence
epicc_common <- most_occurence_strain_clsII[epicc_overlap,1]
```

EpiCC Result

```
epicc_common
```

```
## # A tibble: 8 x 1
##                          Sequence
##                             <chr>
## 1   A_SWINE_KANSAS_A01378027_2017
## 2 A_SWINE_ILLINOIS_A01932036_2017
## 3     A_SWINE_IOWA_A02215038_2017
## 4   A_SWINE_KANSAS_A01378038_2017
## 5     A_SWINE_IOWA_A01104104_2017
## 6     A_SWINE_IOWA_A01672518_2017
## 7     A_SWINE_IOWA_A02215202_2017
## 8 A_SWINE_NEBRASKA_A02216645_2017
```

There are 8 Swine IAV found in both Class I and Class II top EpiCC score list. This means 8 of these strains are having relatively high epitope content to the reference vaccine strain compared to other H1N1 Swine IAV sequences.

Part II

About JMX

JanusMatrix (JMX) or it is also called Janus Immunogenicity Score (JIS) - A web-based tool that incorporated a well-established method for MHC (major histocompatibility complex) binding prediction, with a novel assessment of the potential for T cell receptor (TCR) binding based on similarity with self. This means both good MHC binding and poor self-similarity are required for high immunogenicity, i.e. a robust T effector response (He et. al, 2013).

Aim

In this case, we are not looking for self-similarity epitopes but to identify strains that have the most epitopes coverage found in the swine DNA vaccine sequences of MHC Class I / II. From EpiCC analysis, it can only tells us how much relatedness (top EpiCC score) in terms of epitope content but we would not know what epitope sequences are in the content, and JMX is able to.

Materials

Input to JMX tool: DNA vaccine sequences of MHC ClassI / II were used to query against a set of database that comprised of 72 H1N1 Swine IAV sequences.

Output from JMX tool to process and analyze: 8 HTML outfiles for MHC Class I and II respectively.

Methodology/ Working steps

1. Load JMX MHC Class I data and clean up unnecessary rows and columns

Class I PB1 - Data preview

```
colnames(jmx_data_2_clsI_strain)
```

```
## [1] "Filename" "Sequence" "Segment"  "Epitope"
```

```
head(jmx_data_2_clsI_strain[,2:4])
```

```
## # A tibble: 6 x 3
##                           Sequence   Segment   Epitope
##                              <chr>     <chr>     <chr>
## 1    A_SWINE_KANSAS_A01378019_2017 SEGMENT_2 DTVNRTHQY
## 2  A_SWINE_MICHIGAN_A01259076_2017 SEGMENT_2 DTVNRTHQY
## 3    A_SWINE_KANSAS_A01378027_2017 SEGMENT_2 DTVNRTHQY
## 4      A_SWINE_IOWA_A01672518_2017 SEGMENT_2 DTVNRTHQY
## 5  A_SWINE_NEBRASKA_A01672345_2017 SEGMENT_2 DTVNRTHQY
## 6 A_SWINE_MINNESOTA_A01672344_2017 SEGMENT_2 DTVNRTHQY
```

2. Combine all proteins that have epitope sequence hits and calculate the strain frequency

Class I JMX Data preview

```
colnames(jmx_data_all_clsI_matrix)
```

```
## [1] "Sequence" "Segment"  "Epitope"  "Count"    "Presence"
```

```
head(jmx_data_all_clsI_matrix)
```

```
## # A tibble: 6 x 5
## # Groups:   Sequence, Segment [2]
##                           Sequence   Segment   Epitope Count Presence
##                              <chr>     <chr>     <chr> <int>    <dbl>
## 1 A_SWINE_ARKANSAS_A02218161_2017 SEGMENT_2 DTVNRTHQY     1        1
## 2 A_SWINE_ARKANSAS_A02218161_2017 SEGMENT_4 GMIDGWYGY     2        1
## 3 A_SWINE_ARKANSAS_A02218161_2017 SEGMENT_4 NADTLCIGY     1        1
## 4 A_SWINE_ARKANSAS_A02218161_2017 SEGMENT_4 RIYQILAIY     1        1
## 5 A_SWINE_ARKANSAS_A02218161_2017 SEGMENT_4 SVKNGTYDY     1        1
## 6 A_SWINE_ARKANSAS_A02218161_2017 SEGMENT_4 TSADQQSLY     1        1
```

3. Data visualization - Heat map

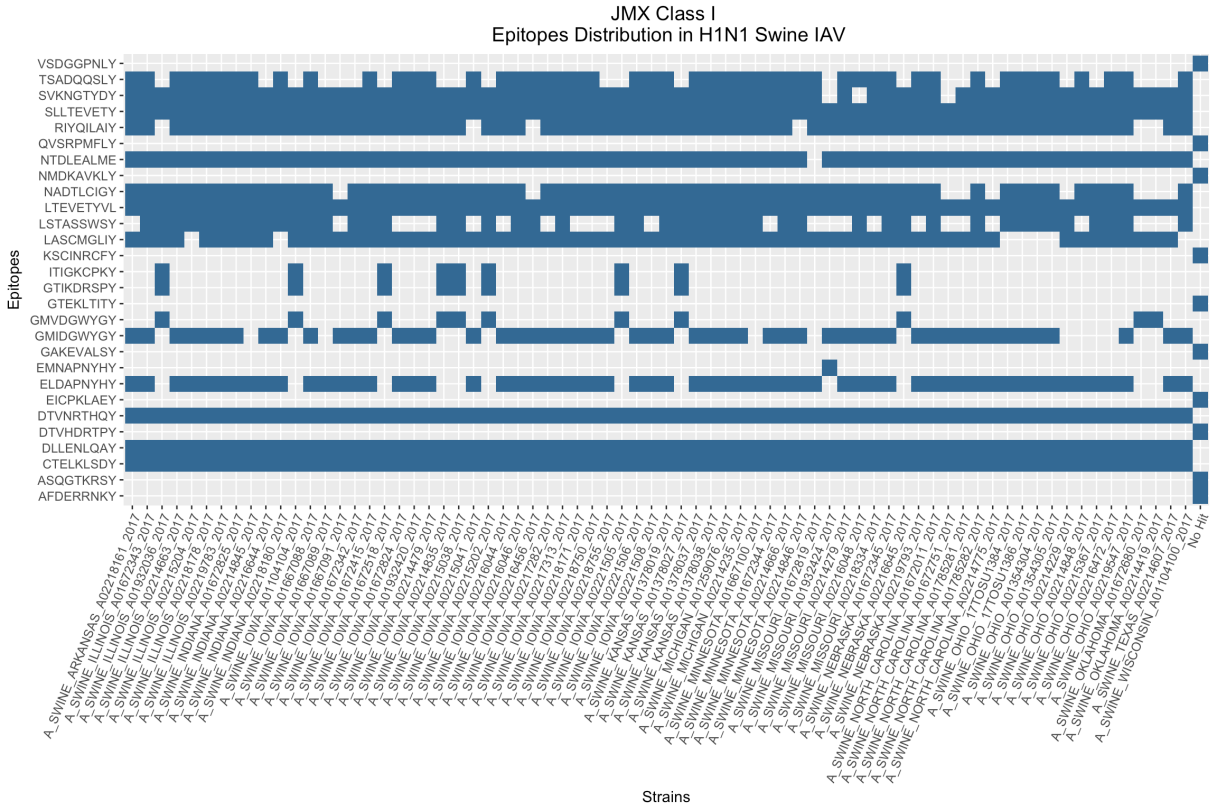

Figure 2.1 | H1N1 Swine IAV strains are plotted against epitopes that found in the DNA vaccine. Blue spots indicate the presence of the epitope, wherease blank spots indicate the opposite. Horizontal view will tell which epitopes are presence and how conserved they are across the strains, while vertical view shows the number of epitopes found in a particular IAV strain.

4. To calculate strain frequency and identify the most occuring strain(s)

Frequency table

```
n_occur
```

```
##                                  Var1 Freq
## 2        A_SWINE_ILLINOIS_A01672343_2017   16
## 4        A_SWINE_ILLINOIS_A02214663_2017   16
## 6        A_SWINE_ILLINOIS_A02218178_2017   16
## 7        A_SWINE_ILLINOIS_A02219783_2017   16
## 8         A_SWINE_INDIANA_A01672825_2017   16
## 13           A_SWINE_IOWA_A01667088_2017   16
## 17           A_SWINE_IOWA_A01672415_2017   16
## 27           A_SWINE_IOWA_A02216046_2017   16
## 30           A_SWINE_IOWA_A02217313_2017   16
## 35           A_SWINE_IOWA_A02221506_2017   16
## 37         A_SWINE_KANSAS_A01378019_2017   16
## 39         A_SWINE_KANSAS_A01378037_2017   16
## 40         A_SWINE_KANSAS_A01378038_2017   16
## 41       A_SWINE_MICHIGAN_A01259076_2017   16
## 42       A_SWINE_MICHIGAN_A02214235_2017   16
## 45      A_SWINE_MINNESOTA_A02214666_2017   16
## 52       A_SWINE_NEBRASKA_A01672345_2017   16
## 55 A_SWINE_NORTH_CAROLINA_A01672011_2017   16
## 58 A_SWINE_NORTH_CAROLINA_A01785282_2017   16
## 68           A_SWINE_OHIO_A02219547_2017   16
## 1        A_SWINE_ARKANSAS_A02218161_2017   15
## 5        A_SWINE_ILLINOIS_A02215204_2017   15
## 10        A_SWINE_INDIANA_A02216644_2017   15
## 11        A_SWINE_INDIANA_A02218180_2017   15
## 12           A_SWINE_IOWA_A01104104_2017   15
## 16           A_SWINE_IOWA_A01672342_2017   15
## 18           A_SWINE_IOWA_A01672518_2017   15
## 19           A_SWINE_IOWA_A01672824_2017   15
## 20           A_SWINE_IOWA_A01932420_2017   15
## 21           A_SWINE_IOWA_A02214479_2017   15
## 22           A_SWINE_IOWA_A02214835_2017   15
## 23           A_SWINE_IOWA_A02215038_2017   15
## 25           A_SWINE_IOWA_A02215202_2017   15
## 26           A_SWINE_IOWA_A02216044_2017   15
## 29           A_SWINE_IOWA_A02217282_2017   15
## 31           A_SWINE_IOWA_A02218171_2017   15
## 32           A_SWINE_IOWA_A02218750_2017   15
## 34           A_SWINE_IOWA_A02221505_2017   15
## 36           A_SWINE_IOWA_A02221508_2017   15
## 38         A_SWINE_KANSAS_A01378027_2017   15
## 44      A_SWINE_MINNESOTA_A01672344_2017   15
## 46      A_SWINE_MINNESOTA_A02214846_2017   15
## 49       A_SWINE_MISSOURI_A02214279_2017   15
## 50       A_SWINE_MISSOURI_A02216048_2017   15
## 51       A_SWINE_MISSOURI_A02218334_2017   15
## 53       A_SWINE_NEBRASKA_A02216645_2017   15
## 54       A_SWINE_NEBRASKA_A02219793_2017   15
## 60          A_SWINE_OHIO_17TOSU1384_2017   15
## 61          A_SWINE_OHIO_17TOSU1386_2017   15
## 62           A_SWINE_OHIO_A01354304_2017   15
## 63           A_SWINE_OHIO_A01354305_2017   15
## 72      A_SWINE_WISCONSIN_A01104100_2017   15
## 3        A_SWINE_ILLINOIS_A01932036_2017   14
## 9         A_SWINE_INDIANA_A02214845_2017   14
## 24           A_SWINE_IOWA_A02215041_2017   14
## 28           A_SWINE_IOWA_A02216456_2017   14
## 33           A_SWINE_IOWA_A02218755_2017   14
## 43      A_SWINE_MINNESOTA_A01667100_2017   14
## 67           A_SWINE_OHIO_A02216472_2017   14
## 14           A_SWINE_IOWA_A01667089_2017   13
## 15           A_SWINE_IOWA_A01667091_2017   13
## 48       A_SWINE_MISSOURI_A01932424_2017   13
## 57 A_SWINE_NORTH_CAROLINA_A01785281_2017   13
## 59 A_SWINE_NORTH_CAROLINA_A02214775_2017   13
## 65           A_SWINE_OHIO_A02214848_2017   13
## 66           A_SWINE_OHIO_A02215367_2017   13
## 71          A_SWINE_TEXAS_A02214607_2017   13
## 47       A_SWINE_MISSOURI_A01672819_2017   12
## 56 A_SWINE_NORTH_CAROLINA_A01672751_2017   12
## 64           A_SWINE_OHIO_A02214229_2017   12
## 69       A_SWINE_OKLAHOMA_A01672680_2017   10
## 70       A_SWINE_OKLAHOMA_A02214419_2017   10
```

```
max(n_occur$Freq)
```

```
## [1] 16
```

```
min(n_occur$Freq)
```

```
## [1] 10
```

To extract top occuring strains

```
jmx_data_all_clsI_freq <- n_occur[n_occur$Freq >= 15,]
nrow(jmx_data_all_clsI_freq)
```

```
## [1] 52
```

```
jmx_data_all_clsI_freq
```

```
##                                  Var1 Freq
## 2        A_SWINE_ILLINOIS_A01672343_2017   16
## 4        A_SWINE_ILLINOIS_A02214663_2017   16
## 6        A_SWINE_ILLINOIS_A02218178_2017   16
## 7        A_SWINE_ILLINOIS_A02219783_2017   16
## 8         A_SWINE_INDIANA_A01672825_2017   16
## 13           A_SWINE_IOWA_A01667088_2017   16
## 17           A_SWINE_IOWA_A01672415_2017   16
## 27           A_SWINE_IOWA_A02216046_2017   16
## 30           A_SWINE_IOWA_A02217313_2017   16
## 35           A_SWINE_IOWA_A02221506_2017   16
## 37         A_SWINE_KANSAS_A01378019_2017   16
## 39         A_SWINE_KANSAS_A01378037_2017   16
## 40         A_SWINE_KANSAS_A01378038_2017   16
## 41       A_SWINE_MICHIGAN_A01259076_2017   16
## 42       A_SWINE_MICHIGAN_A02214235_2017   16
## 45      A_SWINE_MINNESOTA_A02214666_2017   16
## 52       A_SWINE_NEBRASKA_A01672345_2017   16
## 55 A_SWINE_NORTH_CAROLINA_A01672011_2017   16
## 58 A_SWINE_NORTH_CAROLINA_A01785282_2017   16
## 68           A_SWINE_OHIO_A02219547_2017   16
## 1        A_SWINE_ARKANSAS_A02218161_2017   15
## 5        A_SWINE_ILLINOIS_A02215204_2017   15
## 10        A_SWINE_INDIANA_A02216644_2017   15
## 11        A_SWINE_INDIANA_A02218180_2017   15
## 12           A_SWINE_IOWA_A01104104_2017   15
## 16           A_SWINE_IOWA_A01672342_2017   15
## 18           A_SWINE_IOWA_A01672518_2017   15
## 19           A_SWINE_IOWA_A01672824_2017   15
## 20           A_SWINE_IOWA_A01932420_2017   15
## 21           A_SWINE_IOWA_A02214479_2017   15
## 22           A_SWINE_IOWA_A02214835_2017   15
## 23           A_SWINE_IOWA_A02215038_2017   15
## 25           A_SWINE_IOWA_A02215202_2017   15
## 26           A_SWINE_IOWA_A02216044_2017   15
## 29           A_SWINE_IOWA_A02217282_2017   15
## 31           A_SWINE_IOWA_A02218171_2017   15
## 32           A_SWINE_IOWA_A02218750_2017   15
## 34           A_SWINE_IOWA_A02221505_2017   15
## 36           A_SWINE_IOWA_A02221508_2017   15
## 38         A_SWINE_KANSAS_A01378027_2017   15
## 44      A_SWINE_MINNESOTA_A01672344_2017   15
## 46      A_SWINE_MINNESOTA_A02214846_2017   15
## 49       A_SWINE_MISSOURI_A02214279_2017   15
## 50       A_SWINE_MISSOURI_A02216048_2017   15
## 51       A_SWINE_MISSOURI_A02218334_2017   15
## 53       A_SWINE_NEBRASKA_A02216645_2017   15
## 54       A_SWINE_NEBRASKA_A02219793_2017   15
## 60          A_SWINE_OHIO_17TOSU1384_2017   15
## 61          A_SWINE_OHIO_17TOSU1386_2017   15
## 62           A_SWINE_OHIO_A01354304_2017   15
## 63           A_SWINE_OHIO_A01354305_2017   15
## 72      A_SWINE_WISCONSIN_A01104100_2017   15
```

5. JMX Class II Data - Load and clean up unnecessary columns and rows

[Note] There is slightly different in terms of approach and data interpretation compare to JMX Class I due to the length of epiotpes targeted in Class I and Class II are different. Class I epitopes are of the exact length of 9 amino acids, whereas the length of Class II epitopes are between 17-24 amino acids.

Class II PB1 - Data preview

```
head(jmx_data_2_clsII_strain[,2:5])
```

```
## # A tibble: 6 x 4
##                         Sequence   Segment String  Epitopes
##                            <chr>     <chr>  <chr>     <chr>
## 1 A_SWINE_KANSAS_A01378019_2017 SEGMENT_2      1 MMGMFNMLS
## 2 A_SWINE_KANSAS_A01378019_2017 SEGMENT_2      2 RYGFVANFS
## 3 A_SWINE_KANSAS_A01378019_2017 SEGMENT_2      4 MFNMLSTVL
## 4 A_SWINE_KANSAS_A01378019_2017 SEGMENT_2      5 FVANFSMEL
## 5 A_SWINE_KANSAS_A01378019_2017 SEGMENT_2      5 FNMLSTVLG
## 6 A_SWINE_KANSAS_A01378019_2017 SEGMENT_2      8 LSTVLGVSI
```

6. Combine all proteins and calculate frequency of epitope hits for each strain

Class II JMX Data preview

```
colnames(jmx_data_all_clsII_matrix)
```

```
## [1] "Sequence" "Segment"  "Epitopes" "Count"
```

```
head(jmx_data_all_clsII_matrix)
```

```
## # A tibble: 6 x 4
## # Groups:   Sequence, Segment [1]
##                           Sequence   Segment  Epitopes Count
##                              <chr>     <chr>     <chr> <int>
## 1 A_SWINE_ARKANSAS_A02218161_2017 SEGMENT_2 FNMLSTVLG     8
## 2 A_SWINE_ARKANSAS_A02218161_2017 SEGMENT_2 FSMELPSFG     8
## 3 A_SWINE_ARKANSAS_A02218161_2017 SEGMENT_2 FVANFSMEL     8
## 4 A_SWINE_ARKANSAS_A02218161_2017 SEGMENT_2 LSTVLGVSI     8
## 5 A_SWINE_ARKANSAS_A02218161_2017 SEGMENT_2 MFNMLSTVL     8
## 6 A_SWINE_ARKANSAS_A02218161_2017 SEGMENT_2 MMGMFNMLS     8
```

7. Class II - to calculate Class II strain frequency

Frequency table

```
n_occur_clsII
```

```
##                                  Var1 Freq
## 12           A_SWINE_IOWA_A01104104_2017  649
## 18           A_SWINE_IOWA_A01672518_2017  649
## 25           A_SWINE_IOWA_A02215202_2017  649
## 53       A_SWINE_NEBRASKA_A02216645_2017  649
## 23           A_SWINE_IOWA_A02215038_2017  638
## 34           A_SWINE_IOWA_A02221505_2017  608
## 38         A_SWINE_KANSAS_A01378027_2017  600
## 69       A_SWINE_OKLAHOMA_A01672680_2017  563
## 55 A_SWINE_NORTH_CAROLINA_A01672011_2017  562
## 2        A_SWINE_ILLINOIS_A01672343_2017  559
## 4        A_SWINE_ILLINOIS_A02214663_2017  559
## 6        A_SWINE_ILLINOIS_A02218178_2017  559
## 7        A_SWINE_ILLINOIS_A02219783_2017  559
## 9         A_SWINE_INDIANA_A02214845_2017  559
## 13           A_SWINE_IOWA_A01667088_2017  559
## 17           A_SWINE_IOWA_A01672415_2017  559
## 27           A_SWINE_IOWA_A02216046_2017  559
## 30           A_SWINE_IOWA_A02217313_2017  559
## 33           A_SWINE_IOWA_A02218755_2017  559
## 35           A_SWINE_IOWA_A02221506_2017  559
## 41       A_SWINE_MICHIGAN_A01259076_2017  559
## 42       A_SWINE_MICHIGAN_A02214235_2017  559
## 45      A_SWINE_MINNESOTA_A02214666_2017  559
## 46      A_SWINE_MINNESOTA_A02214846_2017  559
## 52       A_SWINE_NEBRASKA_A01672345_2017  559
## 65           A_SWINE_OHIO_A02214848_2017  559
## 68           A_SWINE_OHIO_A02219547_2017  559
## 37         A_SWINE_KANSAS_A01378019_2017  556
## 8         A_SWINE_INDIANA_A01672825_2017  554
## 11        A_SWINE_INDIANA_A02218180_2017  548
## 60          A_SWINE_OHIO_17TOSU1384_2017  548
## 61          A_SWINE_OHIO_17TOSU1386_2017  548
## 62           A_SWINE_OHIO_A01354304_2017  548
## 63           A_SWINE_OHIO_A01354305_2017  548
## 72      A_SWINE_WISCONSIN_A01104100_2017  548
## 24           A_SWINE_IOWA_A02215041_2017  547
## 50       A_SWINE_MISSOURI_A02216048_2017  546
## 56 A_SWINE_NORTH_CAROLINA_A01672751_2017  543
## 57 A_SWINE_NORTH_CAROLINA_A01785281_2017  543
## 3        A_SWINE_ILLINOIS_A01932036_2017  541
## 5        A_SWINE_ILLINOIS_A02215204_2017  541
## 22           A_SWINE_IOWA_A02214835_2017  539
## 70       A_SWINE_OKLAHOMA_A02214419_2017  532
## 58 A_SWINE_NORTH_CAROLINA_A01785282_2017  528
## 43      A_SWINE_MINNESOTA_A01667100_2017  520
## 28           A_SWINE_IOWA_A02216456_2017  519
## 10        A_SWINE_INDIANA_A02216644_2017  518
## 64           A_SWINE_OHIO_A02214229_2017  518
## 66           A_SWINE_OHIO_A02215367_2017  518
## 54       A_SWINE_NEBRASKA_A02219793_2017  511
## 29           A_SWINE_IOWA_A02217282_2017  507
## 39         A_SWINE_KANSAS_A01378037_2017  507
## 40         A_SWINE_KANSAS_A01378038_2017  507
## 67           A_SWINE_OHIO_A02216472_2017  497
## 1        A_SWINE_ARKANSAS_A02218161_2017  480
## 20           A_SWINE_IOWA_A01932420_2017  480
## 31           A_SWINE_IOWA_A02218171_2017  480
## 36           A_SWINE_IOWA_A02221508_2017  480
## 44      A_SWINE_MINNESOTA_A01672344_2017  480
## 26           A_SWINE_IOWA_A02216044_2017  465
## 16           A_SWINE_IOWA_A01672342_2017  462
## 51       A_SWINE_MISSOURI_A02218334_2017  462
## 59 A_SWINE_NORTH_CAROLINA_A02214775_2017  446
## 14           A_SWINE_IOWA_A01667089_2017  440
## 19           A_SWINE_IOWA_A01672824_2017  438
## 21           A_SWINE_IOWA_A02214479_2017  438
## 49       A_SWINE_MISSOURI_A02214279_2017  438
## 15           A_SWINE_IOWA_A01667091_2017  423
## 48       A_SWINE_MISSOURI_A01932424_2017  420
## 71          A_SWINE_TEXAS_A02214607_2017  414
## 32           A_SWINE_IOWA_A02218750_2017  413
## 47       A_SWINE_MISSOURI_A01672819_2017  391
```

```
max(n_occur_clsII$Freq)
```

```
## [1] 649
```

```
min(n_occur_clsII$Freq)
```

```
## [1] 391
```

8. Data visualization of JMX Class II data

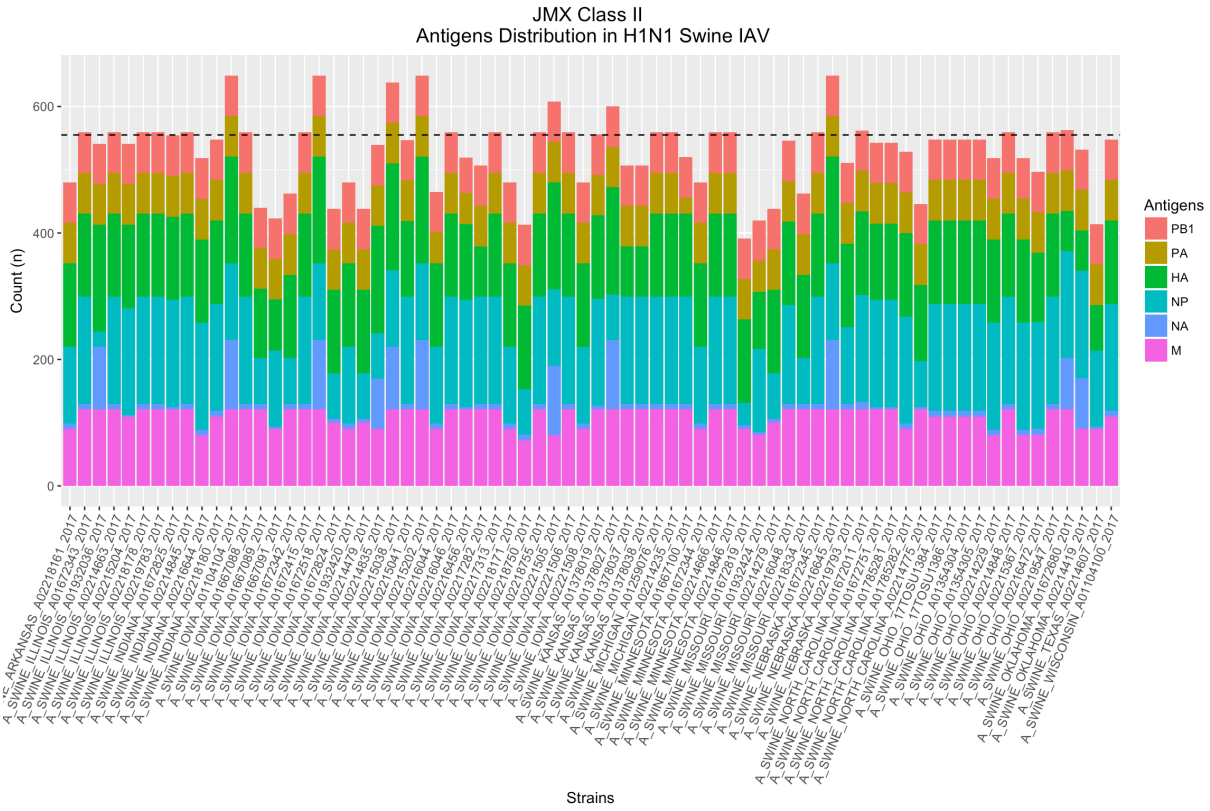

Figure 2.2 | The stacked bar chart shows the total epitopes (in each antigen) found in H1N1 Swine IAV strains. A reference line is drawn across the bar plot and strains that have total frequency equal or above the reference line will be considered.

Class II - To extract strains that meet the cut off value

```
jmx_data_all_clsII_freq <- n_occur_clsII[n_occur_clsII$Freq >= 555,]
nrow(jmx_data_all_clsII_freq)
```

```
## [1] 28
```

```
jmx_data_all_clsII_freq
```

```
##                                  Var1 Freq
## 12           A_SWINE_IOWA_A01104104_2017  649
## 18           A_SWINE_IOWA_A01672518_2017  649
## 25           A_SWINE_IOWA_A02215202_2017  649
## 53       A_SWINE_NEBRASKA_A02216645_2017  649
## 23           A_SWINE_IOWA_A02215038_2017  638
## 34           A_SWINE_IOWA_A02221505_2017  608
## 38         A_SWINE_KANSAS_A01378027_2017  600
## 69       A_SWINE_OKLAHOMA_A01672680_2017  563
## 55 A_SWINE_NORTH_CAROLINA_A01672011_2017  562
## 2        A_SWINE_ILLINOIS_A01672343_2017  559
## 4        A_SWINE_ILLINOIS_A02214663_2017  559
## 6        A_SWINE_ILLINOIS_A02218178_2017  559
## 7        A_SWINE_ILLINOIS_A02219783_2017  559
## 9         A_SWINE_INDIANA_A02214845_2017  559
## 13           A_SWINE_IOWA_A01667088_2017  559
## 17           A_SWINE_IOWA_A01672415_2017  559
## 27           A_SWINE_IOWA_A02216046_2017  559
## 30           A_SWINE_IOWA_A02217313_2017  559
## 33           A_SWINE_IOWA_A02218755_2017  559
## 35           A_SWINE_IOWA_A02221506_2017  559
## 41       A_SWINE_MICHIGAN_A01259076_2017  559
## 42       A_SWINE_MICHIGAN_A02214235_2017  559
## 45      A_SWINE_MINNESOTA_A02214666_2017  559
## 46      A_SWINE_MINNESOTA_A02214846_2017  559
## 52       A_SWINE_NEBRASKA_A01672345_2017  559
## 65           A_SWINE_OHIO_A02214848_2017  559
## 68           A_SWINE_OHIO_A02219547_2017  559
## 37         A_SWINE_KANSAS_A01378019_2017  556
```

9. To find common strains between Class I and Class II JMX data

```
jmx_overlap <- jmx_data_all_clsI_freq$Var1 %in% jmx_data_all_clsII_freq$Var1
jmx_common <- jmx_data_all_clsI_freq[jmx_overlap,]
```

JMX Result

```
jmx_common[,1]
```

```
##  [1] A_SWINE_ILLINOIS_A01672343_2017
##  [2] A_SWINE_ILLINOIS_A02214663_2017
##  [3] A_SWINE_ILLINOIS_A02218178_2017
##  [4] A_SWINE_ILLINOIS_A02219783_2017
##  [5] A_SWINE_IOWA_A01667088_2017
##  [6] A_SWINE_IOWA_A01672415_2017
##  [7] A_SWINE_IOWA_A02216046_2017
##  [8] A_SWINE_IOWA_A02217313_2017
##  [9] A_SWINE_IOWA_A02221506_2017
## [10] A_SWINE_KANSAS_A01378019_2017
## [11] A_SWINE_MICHIGAN_A01259076_2017
## [12] A_SWINE_MICHIGAN_A02214235_2017
## [13] A_SWINE_MINNESOTA_A02214666_2017
## [14] A_SWINE_NEBRASKA_A01672345_2017
## [15] A_SWINE_NORTH_CAROLINA_A01672011_2017
## [16] A_SWINE_OHIO_A02219547_2017
## [17] A_SWINE_IOWA_A01104104_2017
## [18] A_SWINE_IOWA_A01672518_2017
## [19] A_SWINE_IOWA_A02215038_2017
## [20] A_SWINE_IOWA_A02215202_2017
## [21] A_SWINE_IOWA_A02221505_2017
## [22] A_SWINE_KANSAS_A01378027_2017
## [23] A_SWINE_MINNESOTA_A02214846_2017
## [24] A_SWINE_NEBRASKA_A02216645_2017
## 72 Levels: A_SWINE_ARKANSAS_A02218161_2017 ...
```

There are 24 Swine IAV strains overlapped betwee both Class I and Class II JMX data. This means 24 of these strains are having relatively high epitope matches to the DNA vaccine epitopes.

10. To find overlapping strains between EpiCC and JMX results

```
common_both <- epicc_common$Sequence %in% jmx_common$Var1
which(common_both == TRUE)
```

```
## [1] 1 3 5 6 7 8
```

Final result from two approaches - combining EpiCC and JMX

```
epicc_common[common_both,1]
```

```
## # A tibble: 6 x 1
##                          Sequence
##                             <chr>
## 1   A_SWINE_KANSAS_A01378027_2017
## 2     A_SWINE_IOWA_A02215038_2017
## 3     A_SWINE_IOWA_A01104104_2017
## 4     A_SWINE_IOWA_A01672518_2017
## 5     A_SWINE_IOWA_A02215202_2017
## 6 A_SWINE_NEBRASKA_A02216645_2017
```

Conclusion from the analysis: A total of 6 shortlisted H1N1 Swine IAV strains that can be used as challenge strains. Further decision is subjected to collaborator's point of view as there are few more factors (e.g. the antibodies profile) to be considered before reaching a final decision of picking a challenge strain to use in their vaccine study in pigs.