# Supplemental Material to:

## E Jane Homan and Robert D Bremel

## Are cases of mumps in vaccinated patients attributable to mismatches in both vaccine T-cell and B-cell epitopes? An immunoinformatic analysis

Supplemental File S1

## Figure S1.1 Comparison of the PCAA neural network BEPI predictions vs. Bepipred 1.0b

Comparison of the BEPI predictor used in this study to BEPI Pred 1.0b on the HN proteins from 54 different strains of mumps. Bepipred 1.0b output was used as downloaded from the website http://www.cbs.dtu.dk/services/BepiPred/ the predictions used in this paper were transformed to zero mean and unit variance using a Johnson Sb transform. The statics related to the fit is shown below the figure. Using the principal component NN enables us to integrate analyses of B cell linear epitopes alongside MHC binding.
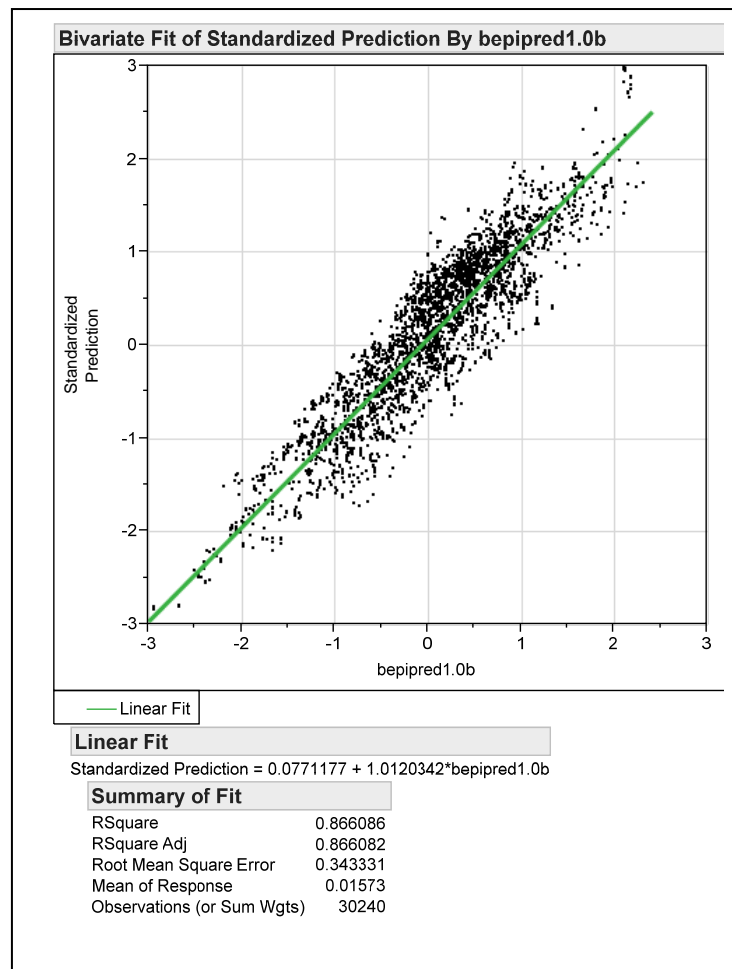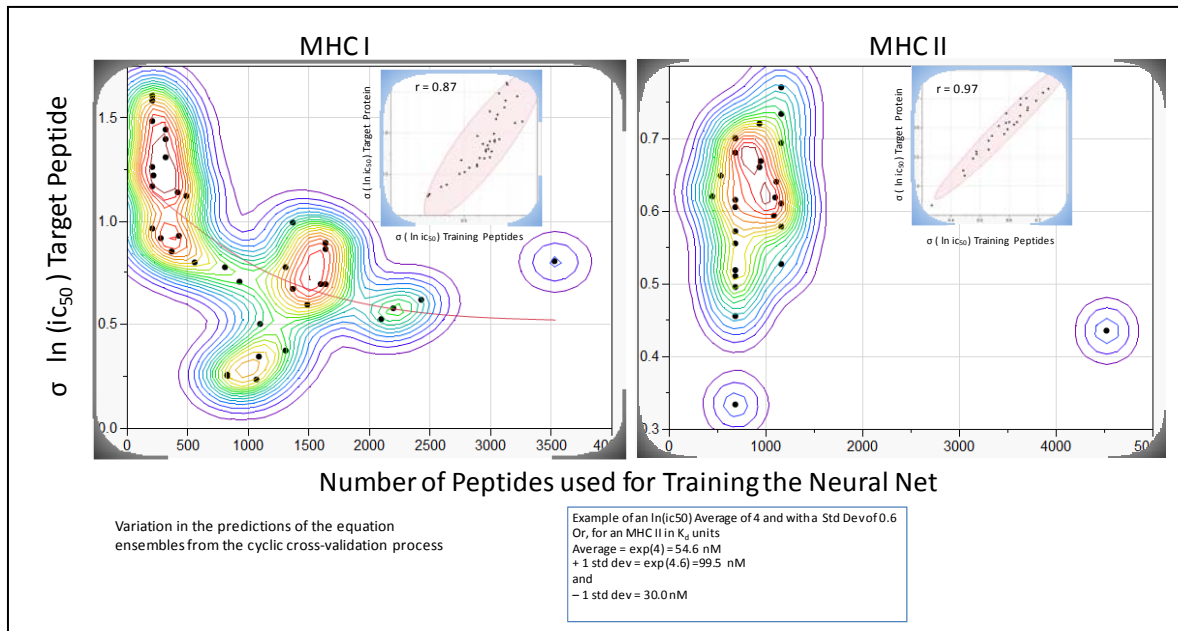


**Bivariate Fit of Standardized Prediction By bepipred1.0b**

—— Linear Fit

**Linear Fit**

Standardized Prediction = 0.0771177 + 1.0120342*bepipred1.0b

**Summary of Fit**

| | |
|---|---|
| RSquare | 0.866086 |
| RSquare Adj | 0.866082 |
| Root Mean Square Error | 0.343331 |
| Mean of Response | 0.01573 |
| Observations (or Sum Wgts) | 30240 |

## Figure S1.2 Variation of Ensembles of Neural Networks predictions MHC binding affinity (ln (ic$_{50}$))

The basic process of using amino acid physical property principal components to predict MHC I and MHC II binding affinity has been described elsewhere (1-3). We have improved on the process described in those publications by use of bootstrap aggregation "bagging" of the training sets to create ensembles of unique prediction equations. The underlying theory behind this process has been described by Brieman (4). Each ensemble member equation produces its

own best estimate of the affinity based on its training and uses a different starting subset of the original training set.  A 5-k-fold cross-validation process is used with each random subset and when the process converges an ensemble member equation is produced that will generate a prediction of binding affinity based on the amino acid sequence of a query peptide.   This process is particularly useful because it makes it possible not only to obtain estimates of the binding affinity but also to provide estimates of the variability of the predictions that are useful to experimentalists.  Each training set has a characteristic variation that seems to depend on the particular combinations of amino acids in the peptides used for the experimental determination of binding affinity.   As seen in the figure the variation is weakly correlated with training set size for MHC I and uncorrelated with size for MHC II.  For MHC II the overall standard deviation in



Variation in the predictions of the equation ensembles from the cyclic cross-validation process

Example of an ln(ic50) Average of 4 and with a Std Dev of 0.6
Or, for an MHC II in $K_d$ units
Average = exp(4) = 54.6 nM
+ 1 std dev = exp(4.6) = 99.5 nM
and
− 1 std dev = 30.0 nM

predictions among ensembles is about 0.6 natural log units.  It is somewhat larger for MHC I alleles. An example of the magnitude of the variation in predictions is shown in the figure.

In addition, with this strategy it is possible to compare the performance of the neural network predictions on both the peptides used for training and on peptides from proteins of interest.  If the networks are performing consistently the correlations should be high.  The insets show that the variation in ensemble predictions is comparable for the training sets and for random target peptides.  Ensembles of 10 equations are used in the predictions.

Panel A. The variation (in standard deviation units) in ensemble predictions  related to training set size for 38 human MHC I alleles.  Panel B.  Variation in predictions for 28 human MHC II alleles related to training set size.  The variation in the predictions are not strongly related to the number of peptides in the training set used to create the ensembles.  The non-parametric contour lines are only to provide visual impression of clusters of somewhat different performance.  The insets show the comparison of the variation ensemble predictions between the training sets and an experimental set of peptides.

Reference List

1. **Bremel RD, Homan EJ.** 2010. An integrated approach to epitope analysis I: Dimensional reduction, visualization and prediction of MHC binding using amino acid principal components and regression approaches. Immunome.Res. **6:**7.
2. **Bremel RD, Homan EJ.** 2010. An integrated approach to epitope analysis II: A system for proteomic-scale prediction of immunological characteristics. Immunome research **6:**8.
3. **Bremel RD, Homan, EJ.** 2013. Recognition of higher order immunologic patterns in proteins: immunologic kernels. PloS One, *In Press*.
4. **Breiman L.** 1996. Bagging predictors. Machine Learning **24:**123-140.

**Figure S3 (Next page): Dendrograms of hierarchical cluster analysis** Hierarchical clusters of 54 different mumps viruses for predicted B-cell contact points and Class I  A and B and Class II DR, DQ and DP MHC binding affinity. Prior to clustering all predicted $\ln(ic_{50})$ predictions for each allele were standardized to zero mean and unit variance within virus.  Hierarchical clustering of the indicated immunological metrics was done using Ward's minimum variance method.  Cluster colors are assigned by software.  Line segments of dendrograms are scaled to distance between viruses for the  indicated metric. Vaccine strains are indicated by open triangles; other strains are wild type.

Hierarchical Clustering allele group=inv_PredBE
Method = Ward

Hierarchical Clustering allele group=HLA-A
Method = Ward

Hierarchical Clustering allele group=HLA-B
Method = Ward
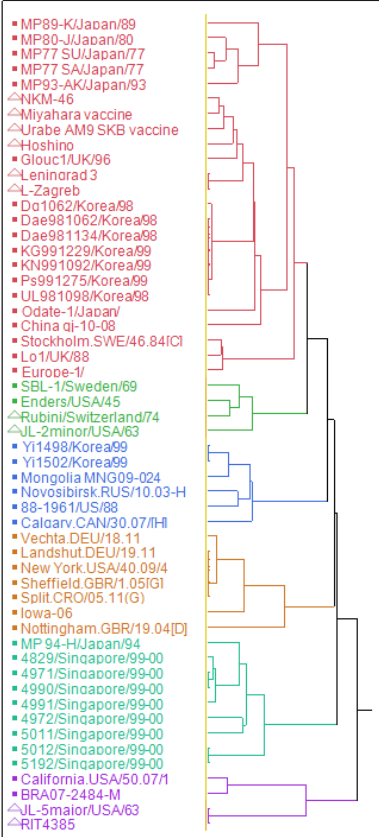
Hierarchical Clustering allele group=HLA-DP
Method = Ward
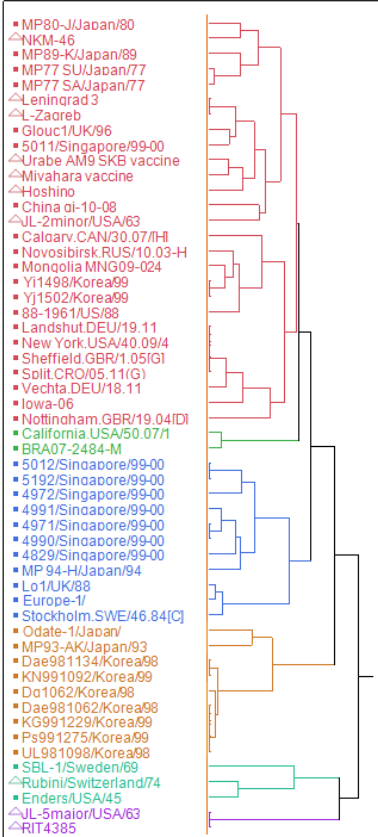
Hierarchical Clustering allele group=HLA-DQ
Method = Ward

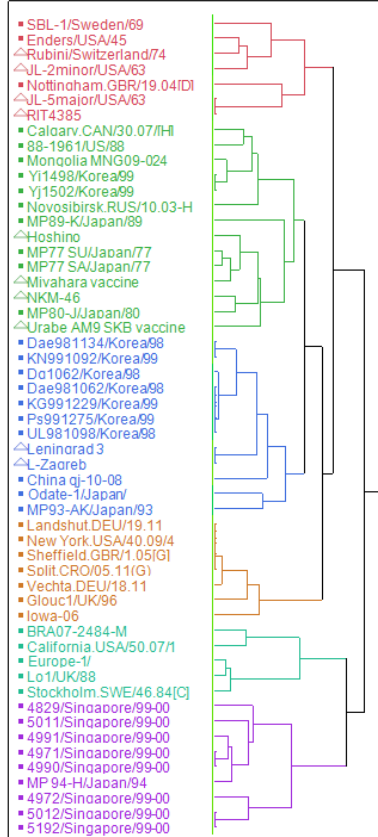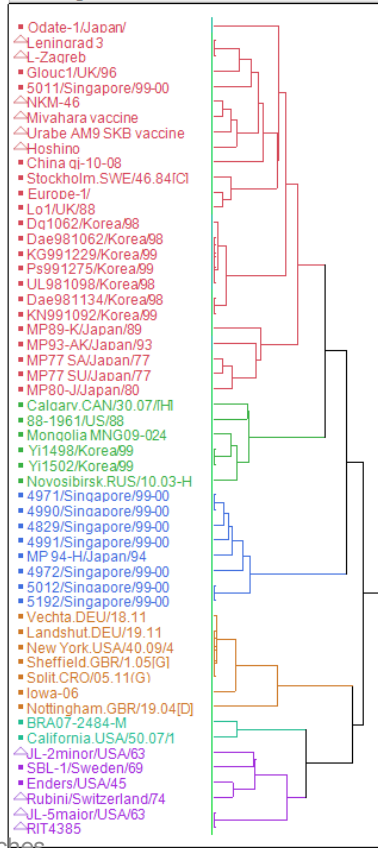Hierarchical Clustering allele group=HLA-DR
Method = Ward

**Figure S4 (Next page): Permuted population plots for five viruses.**
Amino acid positions in HN protein are arrayed N-C on X axis. Red line
shows permuted average predicted MHC-IA and B (37 alleles) binding affinity
by index position of sequential 9-mer peptides. Blue line shows permuted
average predicted MHC-II DRB allele (16 alleles) binding affinity by index
position of sequential 15-mer peptides. Both are plotted in standard deviation
units (Y axis). Orange lines show probability of B-cell binding for an amino
acid centered in each sequential 9-mer peptide. Note that the standardized
B-cell metric has been inverted (multiplied by -1) in order to overlay the
standardized probability onto the same scale as the MHC data. Low numbers
for MHC data represent high binding affinity whereas low numbers equate to
high BEPI contact probability. Bars (Red:MHC-I, Blue:MHC-II) indicate the
top 25% affinity binding. Orange bars indicate high probability B-cell binding.
Background shading shows membrane (green) extramembrane (yellow),
intramembrane (pink) location. The downward orange spike at 269-275
correlates with the experimentally determined principal neutralizing antibody
epitope.

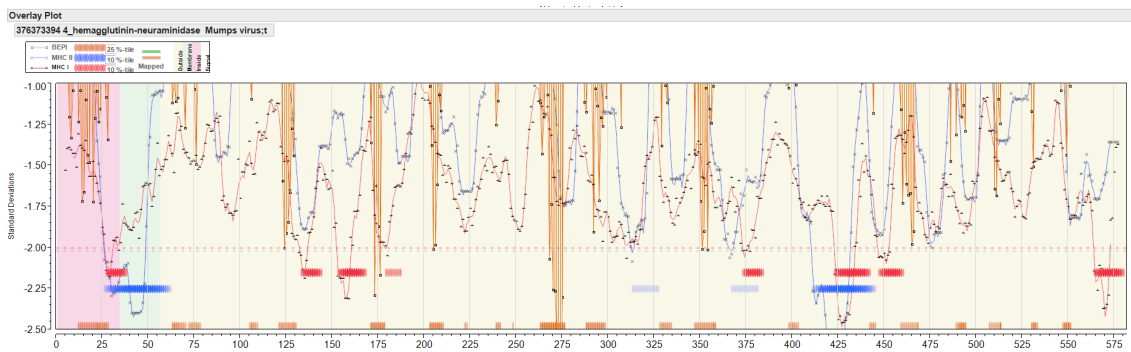Figure S4: Permuted population plots for five viruses. Legend as for Figure 4
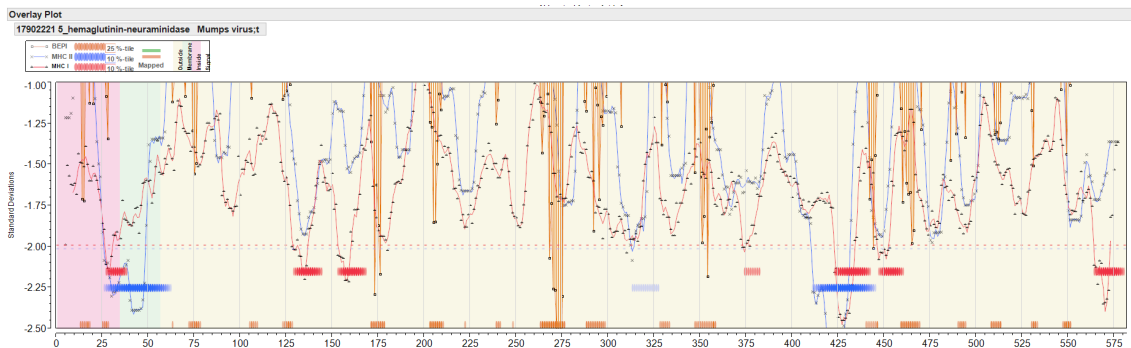
**JL2 Minor**
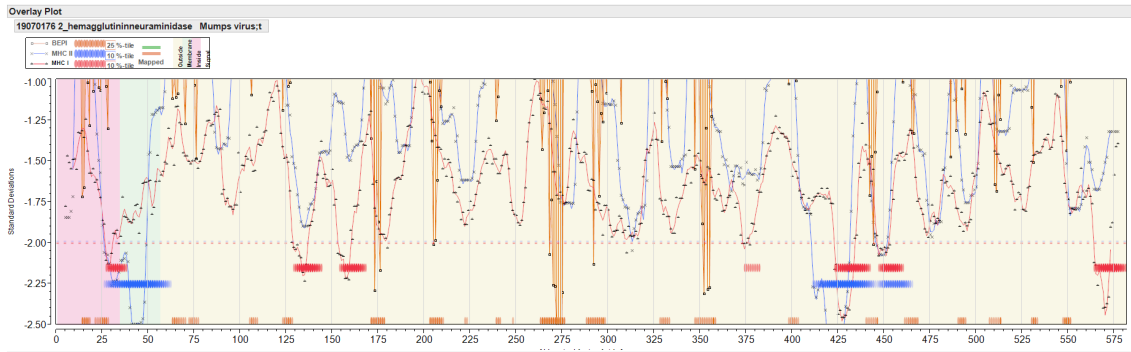


**Lo1/UK/88**
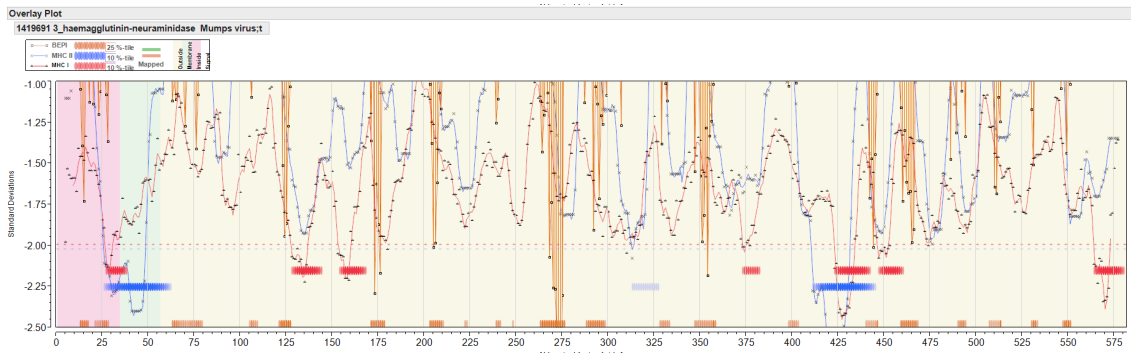


**Calgary.CAN/30.07**



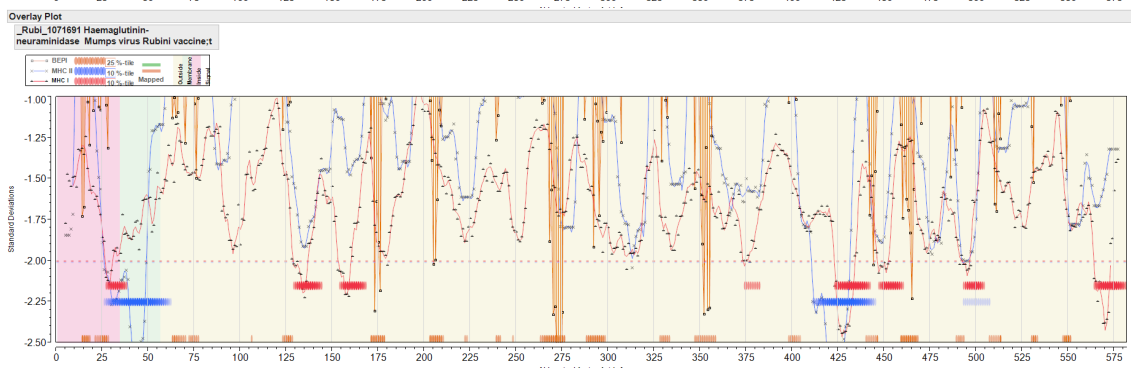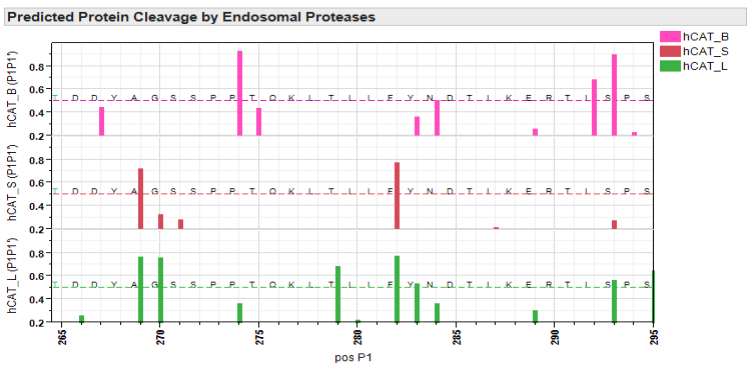**4991/Singapore/99-00**



**Rubini**

**Figure S5 (Next page): Predicted cleavage sites for Cathepsin B, Cathepsin S and Cathepsin L between amino acid positions 265-295**
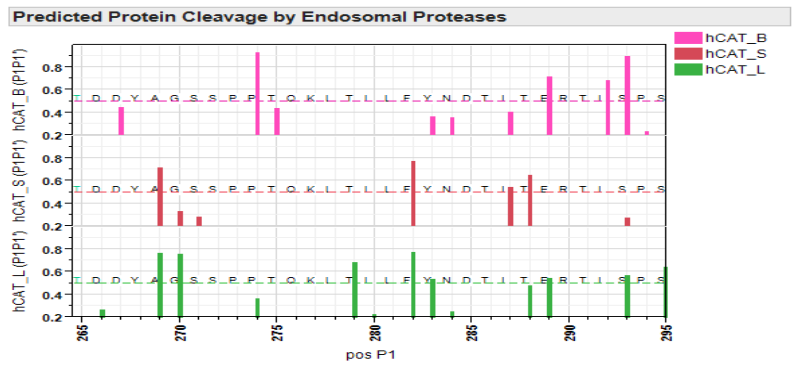X axis shows the amino acid positions from 265-295, one of the regions of high predicted epitope density. Y axis shows the predicted probability of cleavage by one of three different endosomal cathepsins, top to bottom cathespin B (pink), cathepsin S (red) and cathepsin L (green). The bar indicates a cleavage on the C-terminal side of the amino acid at that position. Only probabilities > 0.2 are shown. The dotted lines indicate probability of 0.5.
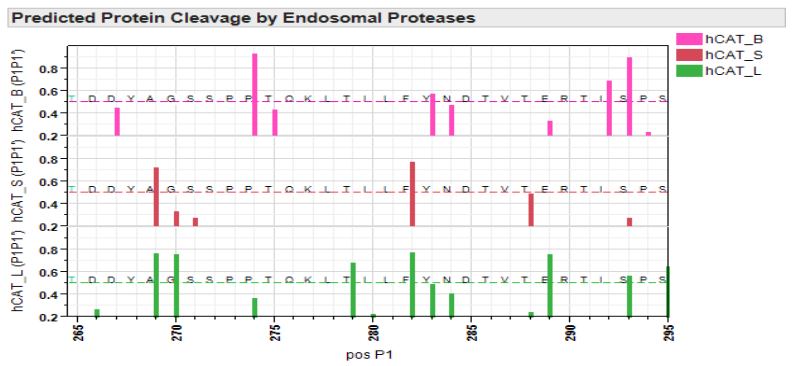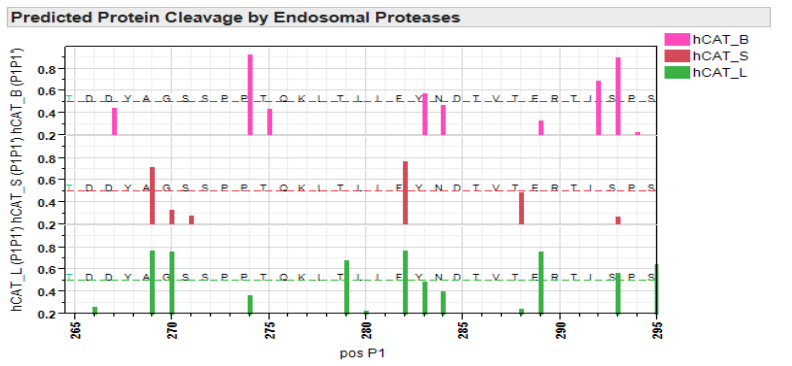
Figure S5:

**JL2 Minor**


Predicted Protein Cleavage by Endosomal Proteases

**Lo1/UK/88**


Predicted Protein Cleavage by Endosomal Proteases

**Calgary.CAN/30.07**


Predicted Protein Cleavage by Endosomal Proteases

**4991/Singapore/99-00**


Predicted Protein Cleavage by Endosomal Proteases

**Rubini**


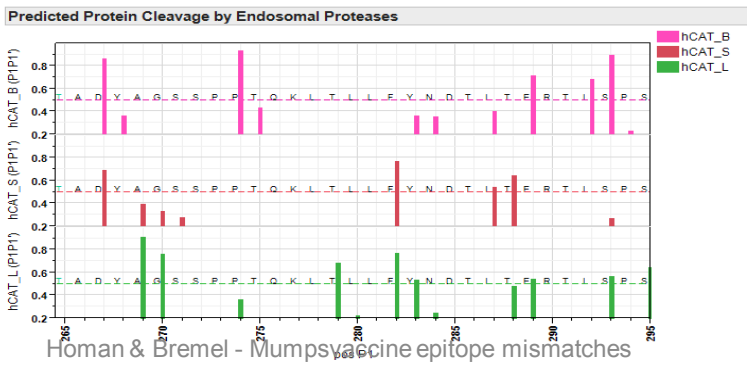Predicted Protein Cleavage by Endosomal Proteases

**Figure S6 (Next page): Predicted cleavage sites for Cathepsin B, Cathepsin S and Cathepsin L between amino acid positions 400-475**
X axis shows the amino acid positions from 400-475. Y axis shows the predicted probability of cleavage by one of three different endosomal cathepsins, top to bottom cathespin B (pink), cathepsin S (red) and cathepsin L (green). The bar indicates a cleavage on the C-terminal side of the amino acid at that position. Only probabilities > 0.2 are shown. The dotted lines indicate probability of 0.5.
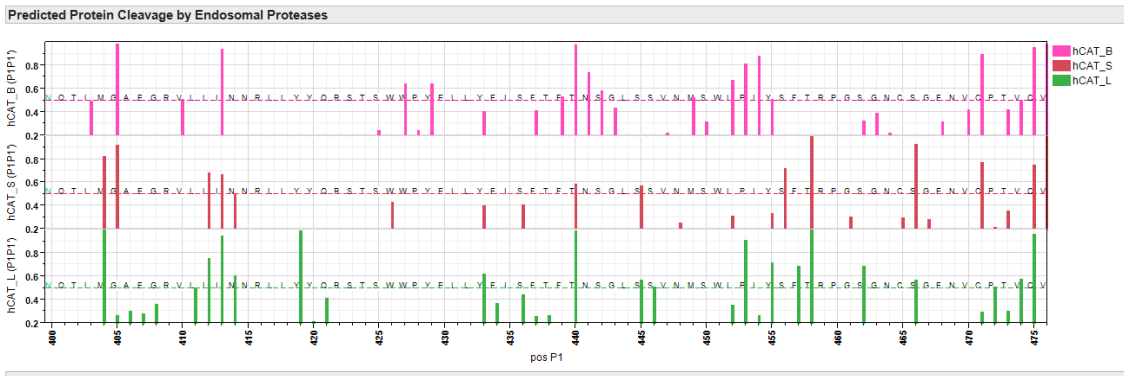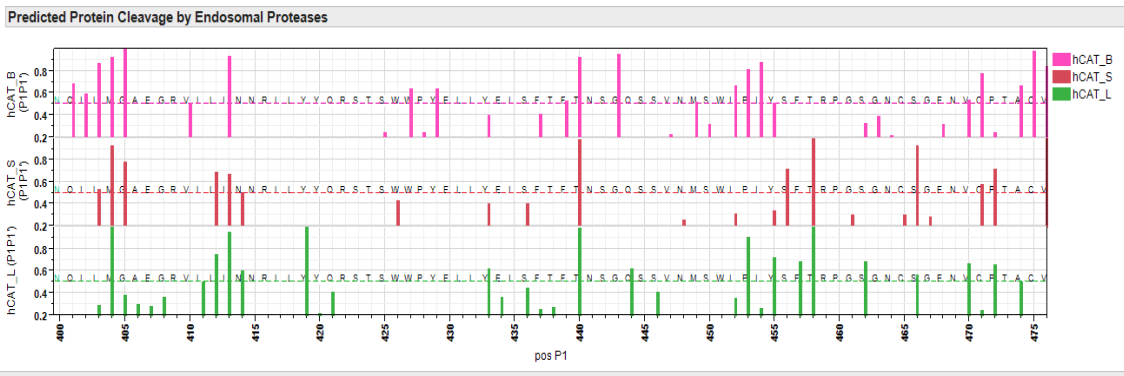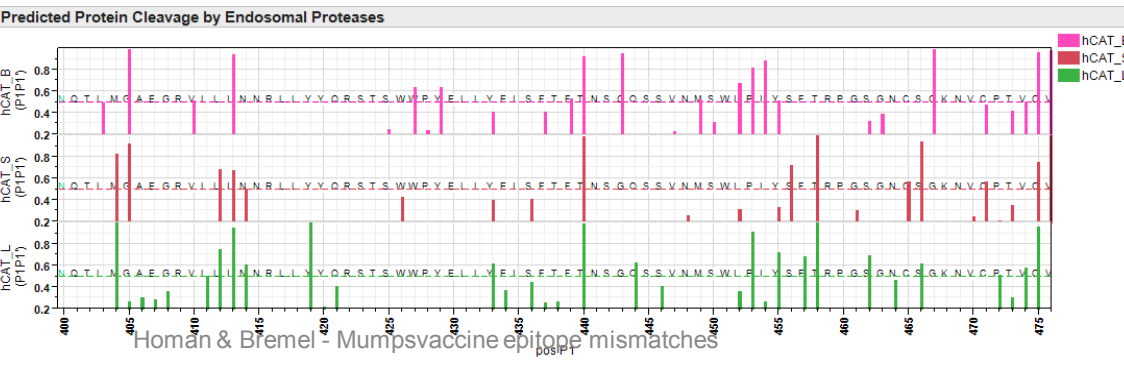
Figure S6:



JL2 Minor
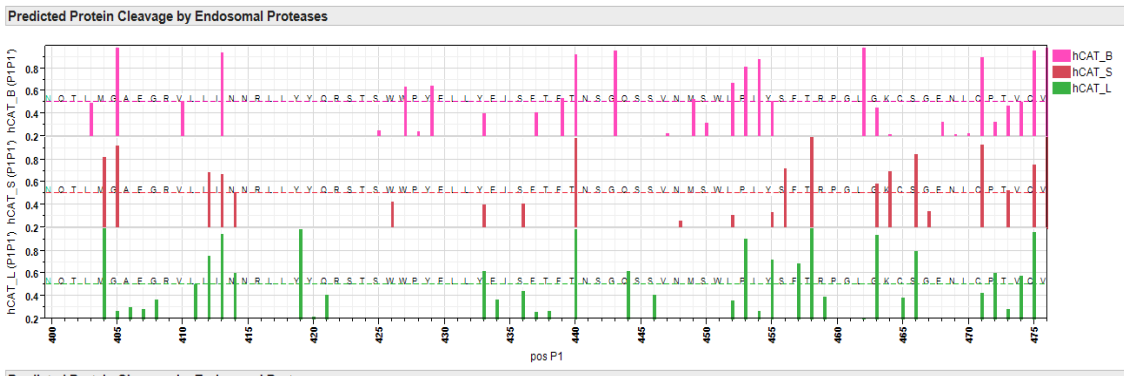
Lo1/UK/88
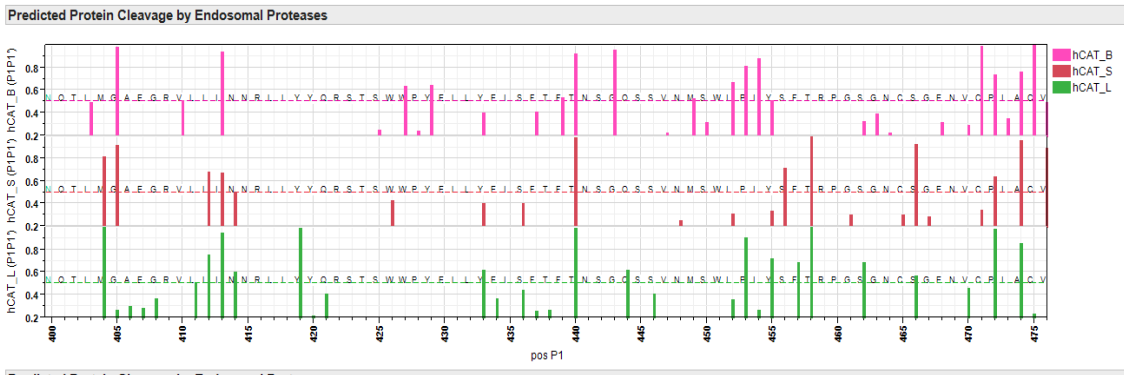
Calgary.CAN/30.07

4991/Singapore/99-00

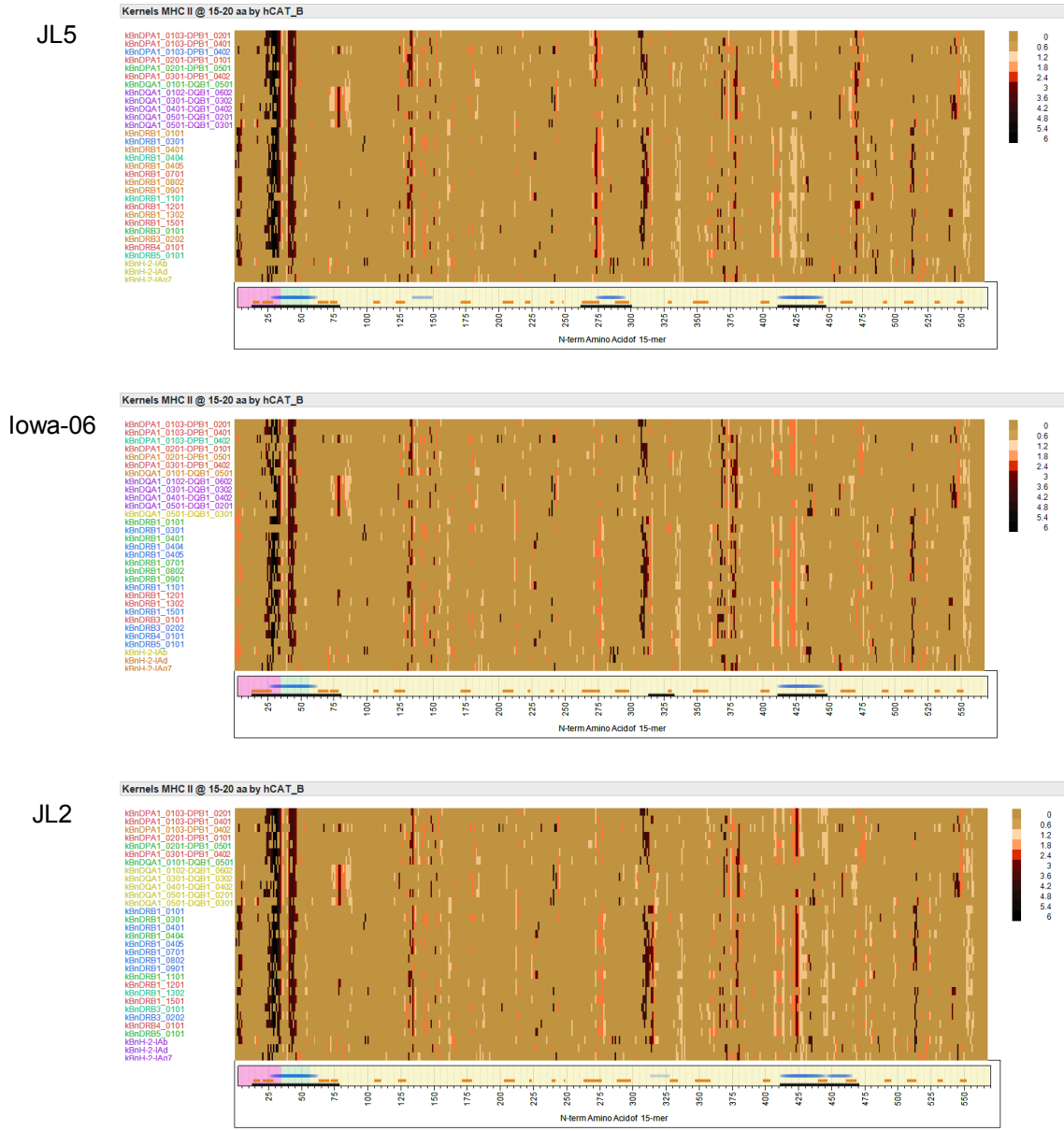Rubini

**Figure S7 (Next page): Predicted high affinity MHC binding peptides following cleavage by cathepsin B, L, and S**

Y Axis shows MHC-II alleles colored by cluster but ranked alphabetically. X axis shows peptide index positions from 260-475. The yellow baseline marginal bar shows the areas of high probability BEPI (orange) and high density MHC-II binding (blue) as in Figure 4. Peptides with high predicted binding affinity (-1 to -3 standard deviations below the mean) for each allele, that are between 15 and 20 amino acids in length and that are predicted to be excised by cathepsin cleavage are colored more intensely. The uniform background coloration indicates that there are no high affinity peptides in that region although there are numerous potential cleavages predicted. The difference in background colors are only to aid in identification of the operative peptidase. Peptides might be cut in any length ranging between 15 and 20 amino acids. They may have several different cut sites. The peptide pixel color intensity is indicative of the number of sites with > 0.5 probability of being cleaved i.e. the darkest colors might be cut in any one of six different positions. As the MHC-II binding groove accommodates multiple lengths of peptide any of the peptides is likely to bind. The arrows indicate the regions where most difference is seen at amino acid index positions 275 and 425.

Figure S7:

## Cathepsin B

JL5



Iowa-06



JL2

# Cathepsin B

Lo1/UK/88
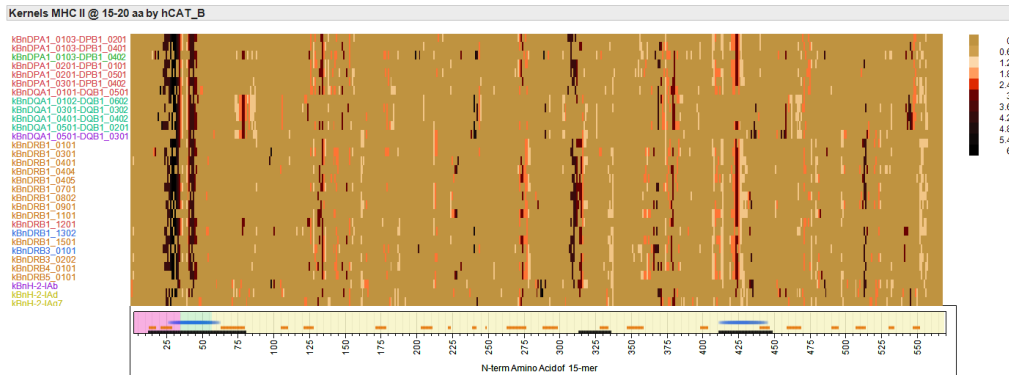


Calgary.CAN/30.07



4991/Singapore/99-00



Rubini

# Cathepsin L

JL5



Iowa-06



JL2

Cathepsin L



Lo1/UK/88

Calgary.CAN/30.07

4991/Singapore/99-00
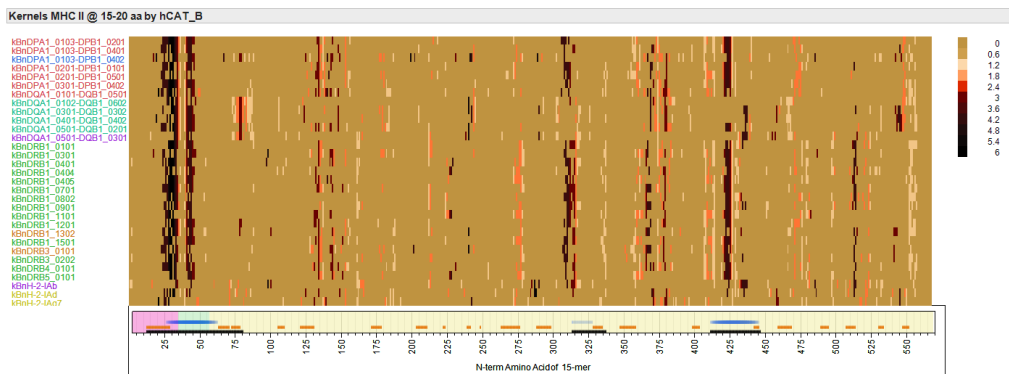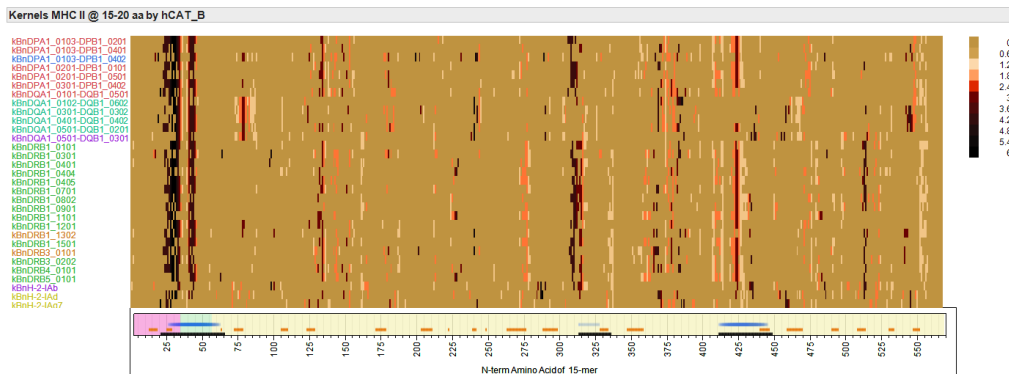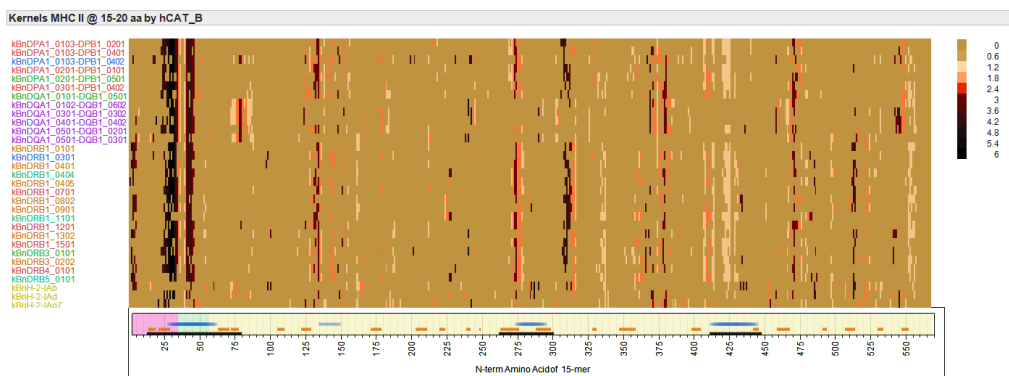
Rubini

Cathepsin S



JL5



Iowa-06



JL2 Minor
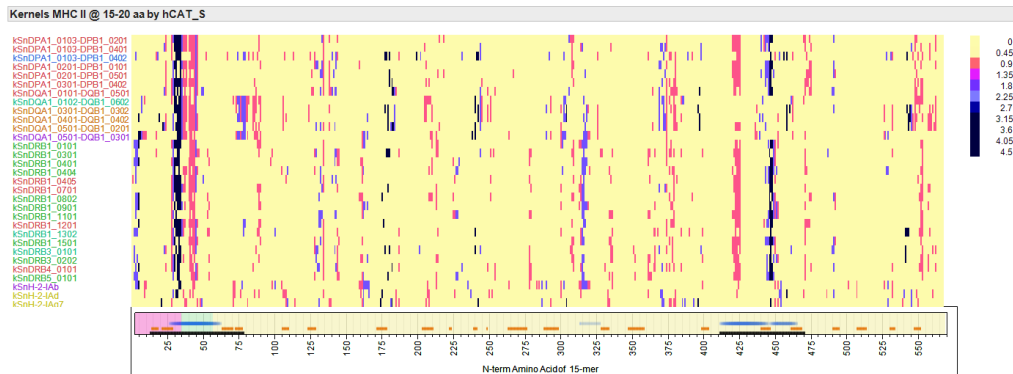
Cathepsin S
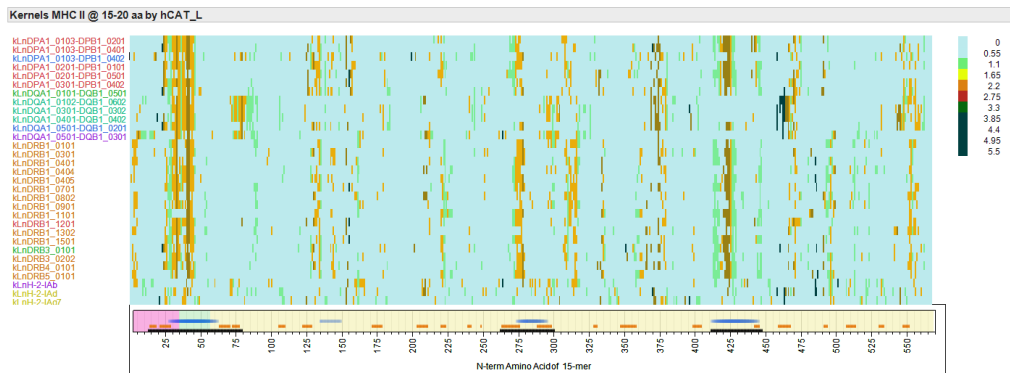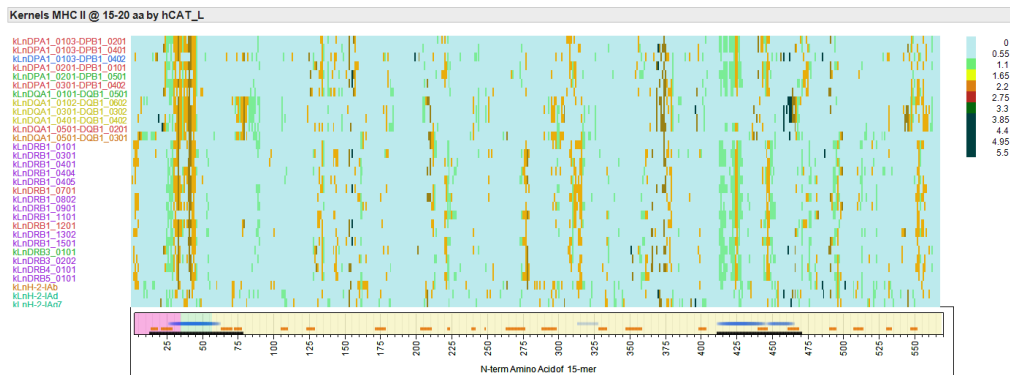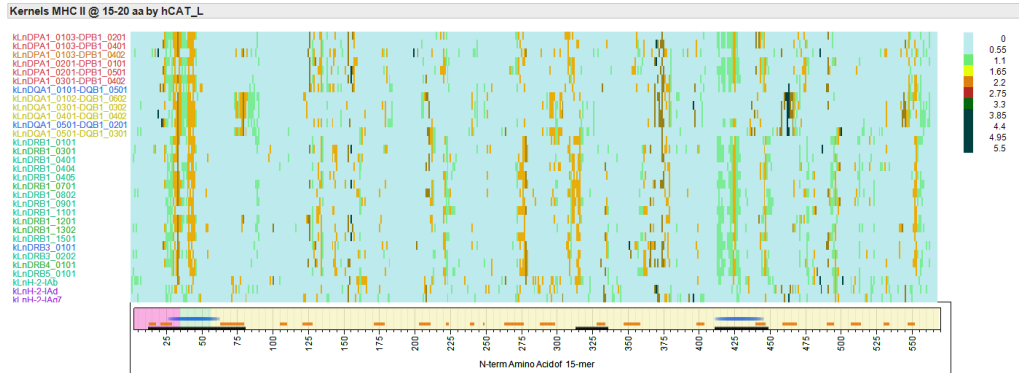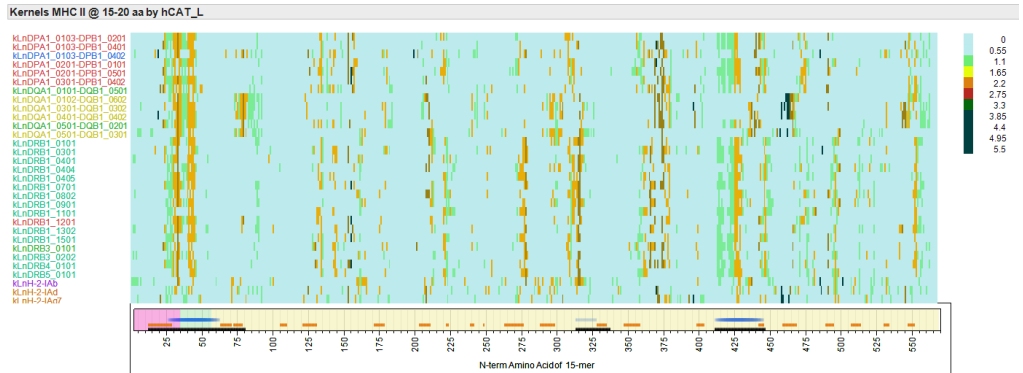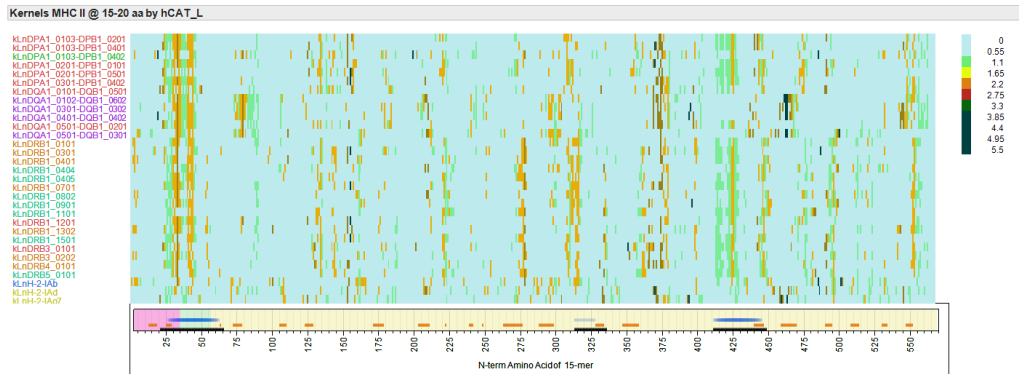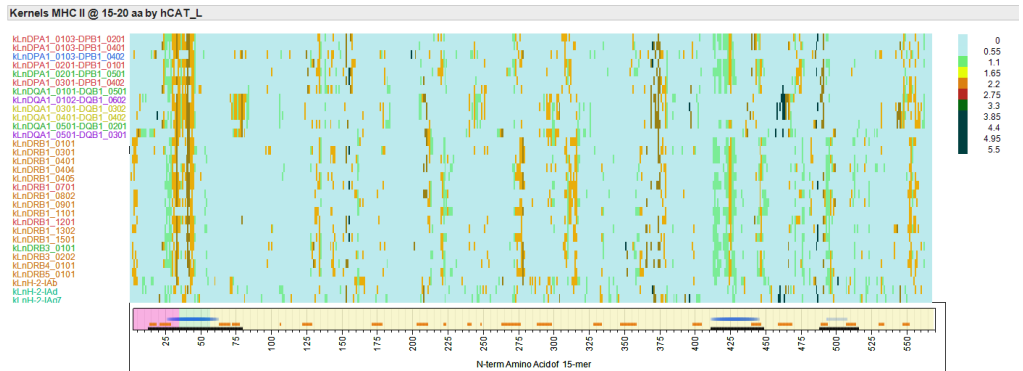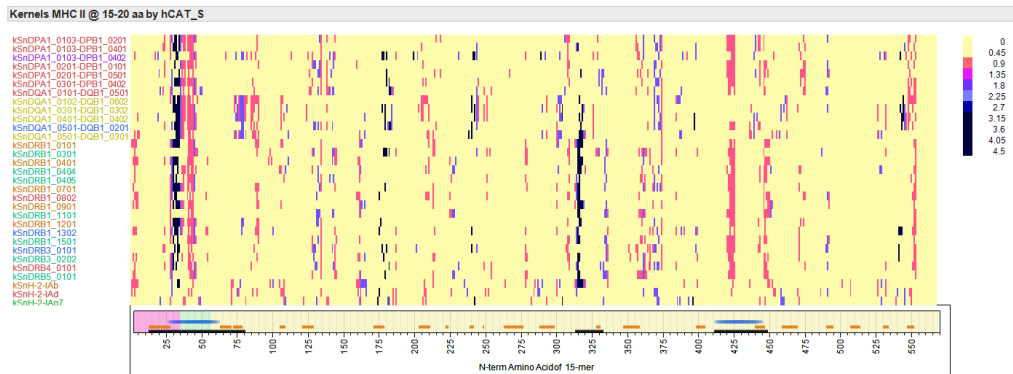


Lo1/UK/88
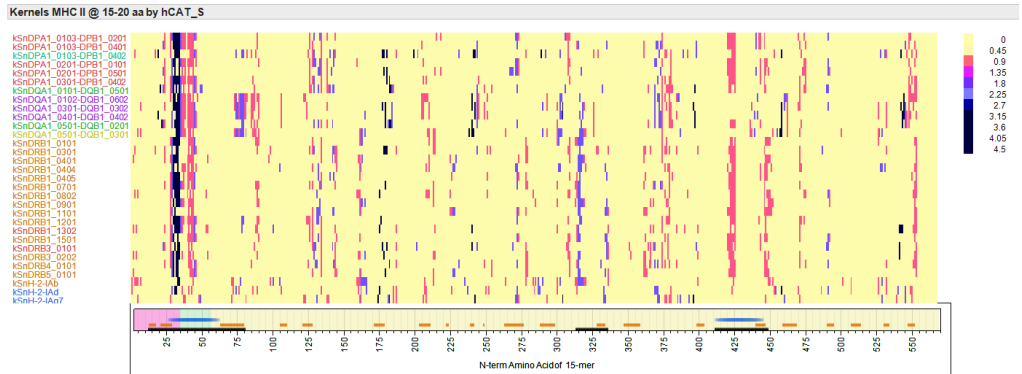
Calgary.CAN/30.07
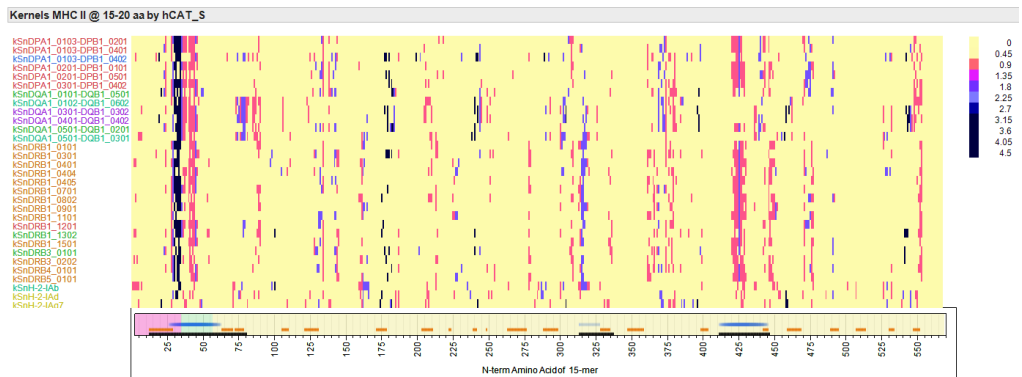
4991/Singapore/99-00

Rubini
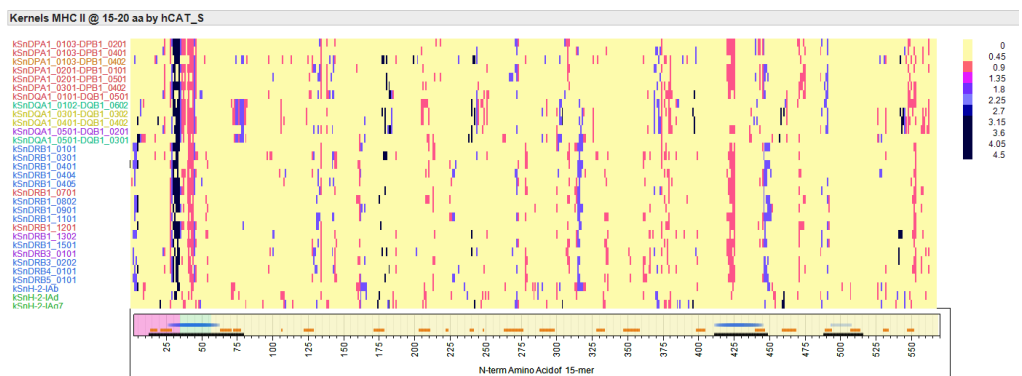
**Figure S8: Sequence alignment of attenuated vaccines in predicted high affinity MHC-II binding regions.**

```
  gi                      277
15077510   JL5          k l i l l f y n d t i t e r t i s p
19070176   JL2          . . t . . . . . . . i k . . . . . .
224482377  RIT4385      . . i . . . . . . . i . . . . . . .
1071691    Rubini       . . t . . . . . . . i . . . . . . .
335347242  Leningrad 3  . . t . . . . . . . v . . . . . . .
55775561   L Zagreb     . . t . . . . . . . v . . . . . . .
213656653  Hoshino      . . t . . . . . . . v . . . . . . .
122989     Miyahara     . . t . . . . . . . v . . . . . . .
14325904   Urabe AM9    . . t . . . . . . . v . . . . . . .
514236698  NKM 46       . . t . . . . . . . v . . . . . . .


           400             440           461
15077510   q t l m g a   n y g q s     g s g h c s g e n v c p i v c v
19070176   . . . . . .   . s . . .   . l . k . . . e . i . . t . . .
224482377  . . . . . .   . y . . .     s . h . . . e . v . . i . . .
1071691    . . . . . .   . s . . .   . s . n . . . k . v . . t . . .
335347242  . . . . . .   . s . . .   . s . n . . . e . v . . t . . .
55775561   . . . . . .   . s . . .   . s . n . . . e . v . . t . . .
213656653  . . . . . .   . s . . .   . s . n . . . e . v . . t . . .
122989     . . . . . .   . s . . .   . s . n . . . e . v . . t . . .
14325904   . . . . . .   . s . . .   . s . k . . . e . v . . t . . .
514236698  . . . . . .   . s . . .   . s . n . . . e . v . . t . . .
```

**Table S1: Source of mumps virus HN proteins analyzed.**

| | Accession number | Strain | SH genotype | GI |
|---|---|---|---|---|
| 1 | AAA74751 | SBL-1/Sweden/69 | A | 332278 |
| **2\*\*#** | **CAA63668** | **Rubini/Switzerland/74** | **A** | **1071691** |
| 3 | CAA63664 | Enders/USA/45 | A | 1071685 |
| **4\*\*#** | **AAL83745** | **JL-2minor/USA/63** | **A** | **19070176** |
| **5\*\*#** | **AAK83223** | **JL-1major/USA/63** | **A** | **15077510** |
| 6# | ACN50032 | RIT4385 | A | 224482377 |
| 7# | BAG84572 | Hoshino | B | 213656653 |
| 8 | BAA77019 | MP 77 SU/Japan/77 | B | 4630936 |
| 9# | P11235 | Miyahara vaccine | B | 122989 |
| 10 | BAA77016 | MP77 SA/Japan/77 | B | 4630932 |
| 11 | BAA77007 | MP89-K/Japan/89 | B | 4630920 |
| 12 | BAA77010 | MP80-J/Japan/80 | B | 4630924 |
| 13# | AAK60072 | Urabe AM9 SKB vaccine | B | 14325904 |
| 14 | AFK80270 | Stockholm.SWE/46.84[C] | **C** | 388890446 |
| **15\*\*** | **CAA67377** | **Lo1/UK/88** | **D** | **1419691** |
| 16 | CAA67378 | Europe-1/ | D | 1419689 |
| 17 | AFJ92903 | Nottingham.GBR/19.04[D] | **D** | 387600834 |
| 18 | AEJ21054 | China qj-10-08 | F | 338832945 |
| 19 | AAG37832 | Glouc1/UK/96 | G | 11545414 |
| 20 | AFO62194 | New York.USA/40.09/4 | G | 397717494 |
| 21 | AFO70355 | Landshut.DEU/19.11 | **G** | 398026084 |
| **22\*\*** | **AEI98831** | **Iowa-06** | **G** | **338784253** |
| 23 | AFO70353 | Vechta.DEU/18.11 | **G** | 398026080 |
| 24 | AFJ41948 | Sheffield.GBR/1.05[G] | **G** | 386832318 |
| 25 | AEY76118 | Split.CRO/05.11(G) | **G** | 373839341 |
| 26 | AAM93227 | Yj1502/Korea/99 | H | 22135493 |
| 27 | AAM93225 | Yj1498/Korea/99 | H | 22135489 |
| 28 | AAL76267 | 88-1961/US/88 | H | 18643332 |
| **29\*\*** | **AFB35810** | **Calgary.CAN/30.07/[H]** | **H** | **376373394** |
| 30 | BAK26761 | Mongolia MNG09-024 | **H** | 333943887 |
| 31 | AAT81471 | Novosibirsk.RUS/10.03-H | **H** | 50812719 |
| 32 | AAP74198 | Dg1062/Korea/98 | I | 32172470 |
| 33 | BAA77013 | MP93-AK/Japan/93 | I | 4630928 |
| 34 | BAA13024 | Odate-1/Japan/ | I | 1468927 |
| 35 | AAM93226 | Ps991275/Korea/99 | I | 22135491 |
| 36 | AAM93224 | KG991229/Korea/99 | I | 22135487 |
| 37 | AAM93223 | KN991092/Korea/99 | I | 22135485 |
| 38 | AAM93222 | Dae981134/Korea/98 | I | 22135483 |
| 39 | AAM93221 | UL981098/Korea/98 | I | 22135481 |
| 40 | AAM93220 | Dae981062/Korea/98 | I | 22135479 |
| 41 | AAL47840 | 4971/Singapore/99-00 | J | 17902233 |
| 42 | AAL47839 | 5192/Singapore/99-00 | J | 17902230 |
| 43 | AAL47838 | 5012/Singapore/99-00 | J | 17902227 |
| 44 | AAL47837 | 5011/Singapore/99-00 | J | 17902224 |
| **45\*\*** | **AAL47836** | **4991/Singapore/99-00** | **J** | **17902221** |
| 46 | AAL47834 | 4972/Singapore/99-00 | J | 17902215 |
| 47 | AAL47833 | 4829/Singapore/99-00 | J | 17902212 |
| 48 | AAL47835 | 4990/Singapore/99-00 | J | 17902218 |
| 49 | BAA76992 | MP 94-H/Japan/94 | J | 4630900 |
| 50 | AFO62149 | California.USA/50.07/1 | K | 397717444 |
| 51 | AEB92303 | BRA07-2484-M | M | 329565933 |
| 52# | AEH42067 | Leningrad 3 | N/A | 335347242 |
| 53# | AY685920 | L-Zagreb | N/A | 55775561 |
| 54# | BAN57616 | NKM-46 | N/A | 514236698 |

\*\* strains selected for further study

\# vaccine strains