Influenza Virus HA Subtype Numbering Conversion Tool and the Identification of Candidate Cross-Reactive Immune Epitopes

Brian J. Reardon, Ph.D.
J. Craig Venter Institute
breardon@jcvi.org
**Background:**

- Interest in comparing amino acid substitutions across influenza A hemagglutinin (HA) subtypes.
- Amino acid equivalence across HA subtypes difficult to determine.
  - Sequence alignment tools unreliable for cross-subtype comparison of HA residues.
    - High sequence variability and differing sequence lengths.
    - Algorithms do not use known structural information.
- Numbering scheme proposed by David Burke and Derek Smith defining amino acid equivalence across HA subtypes (PLoS ONE; 2014).
The Burke-Smith Alignment Method:

Algorithm:
1. Collect representative HA sequences from 18 subtypes of influenza A
2. Trim signal peptide
3. Align sequences with known crystal structural based on structure similarity using PyMOL
4. Align remaining sequences to structure using FUGUE

• Used sequence and structural information utilized to infer amino acid equivalence
Goal at Influenza Research Database (IRD):

Develop a tool incorporating Burke and Smith’s amino acid numbering scheme to facilitate comparison of amino acid substitutions across different HA subtypes.

The HA Subtype Numbering Conversion tool:

- Developed by and freely available at the IRD webpage: fludb.org
- Allows IRD users to computationally convert between HA subtype coordinate systems.
Workflow for the IRD HA Numbering Algorithm

**BLAST**
- Query sequence
  --BLAST against--
  Burke-Smith reference sequences

**Align**
- Query sequence
  --Pairwise alignment--
  BLAST hit

**Map**
- Query sequence
  --mapped to--
  Burke-Smith reference sequence

<table>
<thead>
<tr>
<th>Query H1</th>
<th>Closest BLAST hit</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td>-</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
</tr>
</tbody>
</table>

H1 query position 1 maps to position 4 of BLAST hit (Burke-Smith sequence)

<table>
<thead>
<tr>
<th>Query H1</th>
<th>Closest BLAST hit</th>
<th>H3</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>3</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>4</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>5</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>6</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>7</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>8</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>9</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>10</td>
</tr>
<tr>
<td>-</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>-</td>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td>-</td>
<td>3</td>
<td>13</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>14</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>15</td>
</tr>
</tbody>
</table>

Position 1 of queried H1 sequence maps to position 14 in H3 subtype
Objective:

Use the HA subtype numbering tool in IRD to investigate sequence conservation of putative broadly cross-reactive H1 B-Cell Epitopes across 18 HA subtypes.

General Approach:

1. Obtain H1 B-cell epitope sequences

2. Use HA subtype numbering tool to align H1 B-cell epitopes with the equivalent positions in each HA subtype.

3. Calculate H1 B-cell epitope conservation across HA subtypes
• 93 H1 B-cell epitopes available in IRD
  • Download sequence and coordinates in H1 reference strain

• Navigate to ‘HA Subtype Numbering Conversion’ page in IRD

1. Input (paste) sequence for H1 reference strain
2. Select HA subtypes to identify coordinates of amino acid positions equivalent to the H1 reference strain.
3. Click
## HA Subtype Numbering Conversion Result

The HA Subtype Numbering Conversion tool proposes positions of functional equivalence across different HA subtypes based on the Burke Reference Sequence Alignment. Click here for more details about the algorithm.

### Query Sequence 1

- **Query:** sequence-p|227809830|gb|ACP41105.1| hemagglutinin [Influenza A virus (A/California/04/2009(H1N1))]  
- **Closest Reference sequence:** H1N1pdm (A/California/04/2009)

### Sequence Alignment Result:

<p>| H1N1pdm | H1_PPR34 | H1_1933 | H1post1995 | H2 | H3 | H4 | H5 | H5c221 | H6 | H7N3 | H7N7 | H8 | H9 | H10 | H11 | H12 | H13 | H14 | H15 |</p>
<table>
<thead>
<tr>
<th>--------</th>
<th>----------</th>
<th>---------</th>
<th>------------</th>
<th>----</th>
<th>----</th>
<th>----</th>
<th>----</th>
<th>--------</th>
<th>----</th>
<th>------</th>
<th>------</th>
<th>----</th>
<th>----</th>
<th>-----</th>
<th>-----</th>
<th>-----</th>
<th>-----</th>
<th>-----</th>
<th>-----</th>
<th>-----</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
</tr>
</tbody>
</table>

### Numbering Conversion Result:

- **Query H1N1pdm H1_PPR34 H1_1933 H1post1995 H2 H3 H4 H5 H5c221 H6 H7N3 H7N7 H8 H9 H10 H11 H12 H13 H14 H15**
- **Query:** sequence-p|227809830|gb|ACP41105.1| hemagglutinin [Influenza A virus (A/California/04/2009(H1N1))]  
- **Closest Reference sequence:** H1N1pdm (A/California/04/2009)

### Mapping between positions in epitope and HA sequences

**Query sequence alignment with the closest reference sequence and other HA subtype sequence**

**Mapping between positions in epitope and HA sequences**
Analyzing H1 B-Cell Epitope Conservation Across 18 HA Subtypes

Analysis:

- Organize H1 B-cell epitope and HA subtype sequence data into a spreadsheet so equivalent amino acids are located in the same column
- Count number of matching amino acids in each HA subtype sequence matching epitope

**E.g.,** Conservation of 15 amino acid H1 epitope, SF355 in H2 subtype

<table>
<thead>
<tr>
<th>Amino Acid Position</th>
<th>508</th>
<th>509</th>
<th>510</th>
<th>511</th>
<th>512</th>
<th>513</th>
<th>514</th>
<th>515</th>
<th>516</th>
<th>517</th>
<th>518</th>
<th>519</th>
<th>520</th>
<th>521</th>
<th>522</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1 Epitope: SF355</td>
<td>E</td>
<td>E</td>
<td>A</td>
<td>K</td>
<td>L</td>
<td>N</td>
<td>R</td>
<td>E</td>
<td>E</td>
<td>I</td>
<td>D</td>
<td>G</td>
<td>V</td>
<td>K</td>
<td>L</td>
</tr>
<tr>
<td>HA subtype: H2</td>
<td>E</td>
<td>E</td>
<td>S</td>
<td>K</td>
<td>L</td>
<td>N</td>
<td>R</td>
<td>N</td>
<td>E</td>
<td>I</td>
<td>K</td>
<td>G</td>
<td>V</td>
<td>K</td>
<td>L</td>
</tr>
</tbody>
</table>

- 12 out of 15 amino acids in SF355 match with the corresponding position in the H2 subtype sequence; calculate percent matching amino acids

\[
\text{Percent amino acid identity (per epitope)} = \frac{\# \text{ matches between epitope and HA subtype}}{\text{epitope length}} = \frac{12}{15} \times 100 = 80\%
\]

- For every epitope, calculate percent identity in each HA subtype sequence
Results: Conservation of each Epitope Across All HA Subtypes

H1 B-cell epitope percent amino acid identity in HA subtypes: averaged across all HA subtypes

- No obvious correlation between epitope conservation and secondary structure
- Epitopes in HA2 stem region highly conserved across HA subtypes; particularly near fusion peptide
- Lower overall conservation in HA1 head region, but several epitopes highly conserved across subtypes
Results: Average Epitope Conservation in Each HA Subtype

For each HA subtype, average percent amino acid identity in all 93 H1 epitopes

- H1 B-cell Epitope conservation not equal in all HA subtypes
- H1 epitopes more highly conserved in phylogenetic group 1 (of which H1 is a member)
Conclusions:

**HA Subtype Numbering Conversion tool:**
- Identification and comparison of functionally equivalent amino acids across HA subtypes straightforward using *HA Subtype Numbering Conversion* tool in IRD

**Analysis of H1 B-cell Epitope Conservation across HA subtypes**
- H1 epitopes more conserved in HA subtypes in phylogenetic group 1.
- High epitope conservation across HA subtypes in stem region.
- Lower overall conservation in head region, but several epitopes highly conserved across subtypes
  - Including epitope (amino acids 252-270) adjacent to sialic acid binding site

**Next Steps:**
- Integrate this numbering conversion scheme into other existing tools in IRD
- Extend analysis to include HA sequence data from all strains
Acknowledgements:

J. Craig Venter Institute
Alexandra Lee
Brian Aevermann
Yun Zhang
Brett Pickett
Richard Scheuermann (PI)

Northrop Grumman
Sherry He

University of Auckland
Catherine Macken

University of Cambridge
David Burke
Derek Smith

Thank you

To learn more about the Influenza Research Database, go to fludb.org or stop by our table in the Commonwealth Ballroom (Squires)