Librator, a platform for optimized sequence editing, design, and expression of influenza virus proteins

Lei Li[1], Olivia Stovicek[1], Jenna J. Guthmiller[1], Siriruk Changrob[1], Yanbin Fu[1], Haley L. Dugan[2], Christopher T. Stamper[2], Nai-Ying Zheng[1], Min Huang[1], Patrick C. Wilson[1,2,3,*]

Affiliations:
[1]Section of Rheumatology, Department of Medicine, University of Chicago, Chicago, IL 60637, USA.
[2]Committee on Immunology, University of Chicago, Chicago, IL 60637, USA.
[3]Lead Contact.
*Correspondence: wilsonp@uchicago.edu (P.C.W.)

Abstract

Artificial mutagenesis and chimeric/mosaic protein engineering have laid the foundation for antigenic characterization[1] and universal vaccine design[2-4] for influenza viruses. However, many methods used for influenza research and vaccine development require sequence editing and protein expression, limiting their applicability and the progress of related research to specialists. Rapid tools allowing even novice influenza researchers to properly analyze and visualize influenza protein sequences with accurate nomenclature are needed to expand the research field. To address this need, we developed Librator, a system for analyzing and designing protein sequences of influenza virus Hemagglutinin (HA) and Neuraminidase (NA). With Librator's graphical user interface (GUI) and built-in sequence editing functions, biologists can easily analyze influenza sequences and phylogenies, automatically port sequences to visualize structures, then readily mutate target residues and design sequences for antigen probes and chimeric/mosaic proteins efficiently and accurately. This system provides optimized fragment design for Gibson Assembly[5] of HA and NA expression constructs based on peptide conservation of all historical HA and NA sequences, ensuring fragments are reusable and compatible, allowing for significant reagent savings. Use of Librator will significantly facilitate influenza research and vaccine antigen design.

Main

Influenza is considered to be the next major threat for a devastating pandemic. Vaccination has been proven an effective approach to prevent infection and global spreading of influenza viruses[6]. However, due to frequent mutations in the influenza virus genome, and particularly alterations to the HA and NA surface proteins, influenza vaccine protection is short-lived or mostly ineffective when mismatches have been observed for several flu seasons[7, 8]. In this context, developing universal influenza vaccine candidates that induce broadly reactive immunity and particularly antibodies against conserved epitopes of influenza virus surface proteins is an important direction of research[4, 9-12]. Use of artificial mutagenesis and design of antigen probes and chimeric/mosaic proteins are crucial steps in vaccine development and related research. However, current workflow for these tasks faces several major challenges that make it expensive, fallible and time-consuming. First, there are multiple residue numbering systems for HA protein sequences that have been commonly used in the literature, protocols and the research community[13, 14]. They are Coding Sequence (CDS) position, crystal structure-based H1/H3 numbering[15-17], and Burke and Smith HA numbering. For a given sequence, CDS position usually counts from the start of the CDS,

methionine, and therefore can cover all amino acids; structure-based H1/H3 numbering and Burke and Smith HA numbering determine residue numbers according to template mapping using different templates (see methods section for details). Thus, biologists must put significant effort into identifying the correct residues to avoid errors. Moreover, nucleotide and amino acid sequences are difficult to read, and inefficient and fallible to edit manually. Further, there are no comprehensive tools to develop individual influenza sequence databases and readily compare varied influenza sequences and phylogenies, or to immediately port annotated sequences for visualization of color-coded antigenic-regions, mutations, or epitopes on representative HA and NA structures using structure analysis software such as PyMol and UCSF Chimera. Finally, efficient and automated cloning of HA and NA protein variants to scale can become expensive and is error-prone if done by hand. Gibson Assembly can assemble multiple linear DNA fragments for protein cloning and expression and has been extensively used in molecular biology, and is superior among all assembly methods because of the enormous savings of time and human labor with its easy one-tube reactions. Automated Gibson fragment prediction and databasing of fragments conserved between similar HA and NA protein expression constructs would allow accurate and cost-effective production of variant influenza protein libraries for many applications. In conclusion, these broad challenges in current influenza immunology/virology studies limit the efficiency and breadth of research. Improving the accuracy and economy of these processes will significantly expand related studies both in seasoned influenza laboratories and for novice laboratories interested in applying innovative approaches to the study of influenza.

Here, we present a computational tool called "Librator" for influenza HA and NA sequence analysis, editing, and cost-effective cloning and vector design for HA and NA protein expression. Librator is an integrated graphical processing platform for influenza sequences. Librator seamlessly connects nucleotide sequences (from public sequence databases) and lab work (e.g. Gibson cloning, protein expression), and it contains a variety of functions to facilitate management, analysis, editing and accessing influenza sequences, to improve the efficiency of sequence design and expression (Figure 1A). This software entrusts all error-prone sequence editing and data processing tasks to the background algorithm, so that users are able to design their sequences with a few clicks on the GUI. With Librator users can complete all sequence design-related operations graphically in an integrated system, avoiding difficult-to-read raw sequences or switching between different software applications.

Multiple common functions were built-in to Librator to help users analyze influenza sequences more efficiently. First, an HA numbering aligner was integrated into Librator's sequence viewer and editor. This function aligns given HA sequences to crystal structures of a classic H1 (PDB ID: 4JTV) and a classic H3 template (PDB ID: 4HMG) to identify corresponding H1/H3 numberings for each residue. Known common antigenic sites and epitopes are automatically labelled based on this numbering; for example antibody binding sites (ABS), receptor binding sites (RBS), and the H1 (Ca1, Ca2, Cb, Sa, Sb, Stalk) and H3 (A, B, C, D, E Stalk) antigenic sites are color-coded on H1/H3 numbering rulers (Figure 1B). Notably, epitope definitions in Librator are highly customizable, allowing users to annotate HA sequences according to their specific research interest and focus (Figure S1A). Since glycosylation on the HA protein was reported to have an important impact on antigenic drift[18, 19], an "N-X-S/T" pattern that indicates potential N-linked glycosylation sites are also highlighted. This viewer is also capable of displaying fully annotated multiple sequence alignments with two informative modes, original sequence mode and template mode,

enabling convenient investigation of evolutionary sequence patterns and mutations (Figure 1C). For example, biologists characterizing escape mutant sequences induced by selective pressure with antibodies or sera can align the HA or NA sequences and immediately visualize and export graphics of regions that were mutated. To help users quickly infer phylogenetic relationships among a group of sequences, Librator also allows users to generate and visualize maximal likelihood trees from either nucleotide sequences or peptide sequences (Figure 1D). Furthermore, powered by WebLogo, Librator allows users to access nucleotide and peptide conservation among groups of sequences[20] (Figure 1E). By automatically porting amino acid sequences and labelling instructions to PyMOL[21] or UCSF Chimera[22], Librator allows users to visualize peptides on 3D structures of HA proteins with color-annotated amino acids according to either peptide conservation score (Figure 1E) or all antigenic regions and user-defined sequence labels (Figure 1F). Librator uses an H1 structure (PDB ID: 4JTV) for visualization of all Group 1 HA structures and a H3 structure (PDB ID: 4HMG) for visualization of all Group 2 HA structures[23-25]. For example, with a single button-click, users can immediately evaluate whether an escape mutation is predicted to alter a surface amino acid or occurs deeper in the structure potentially driving conformational changes. In addition, a function was also developed that allows users to identify potential key residues between two groups of sequences by ranking residues by their amino acid difference. For example, by comparing pre-1994 and post-1994 human H1N1 seasonal viruses, Librator highlighted the importance of a deletion "Δ130," which has been validated by experiments[26], by a high ranking score. This tool helps users zero in on important sequence elements driving influenza evolution. Lastly, an independent viewer for users to easily access the Burke and Smith HA numbering scheme proposed by Burke et al. was implemented since it has also been commonly used in the Influenza community[13](Figure S1B). It should be noted, however, that all functions in Librator, including the alignment viewer, sequence editing, and sequence designing, were based on structure-based numbering systems.

To improve the efficiency and accuracy of mutagenesis and sequence editing, we developed multiple functions to help users to design and edit their Influenza sequences. With the help of the HA numbering aligner, users can easily locate target residues and mutate them by simply typing a mutation code using whichever numbering system they prefer. For example, for an H3 sequence (A/England/80740425/2018), typing "Y177M" in CDS position input will mutate the 177th residue of the CDS from Tyrosine (Y) to Methionine (M). This is equivalent to typing "Y164M" in H1 numbering HA1 input or typing "Y161M" in H3 numbering HA1 input (Figure 2A). By translating between the various numbering schema, Librator avoids confusion and mistakes that are common in analyzing influenza sequence data. For NA sequences, only CDS position input is available since it is the only numbering system for NA sequences. To avoid mistakes, Librator validates the original amino acid in the mutation code to make sure it matches the amino acid in the raw sequence in the numbering system used. Expression of influenza HA soluble proteins for experimental purposes is an important tool for characterizing influenza immunity or monoclonal antibody specificity. Building on this mutagenesis function, we also developed a function to design HA expression constructs for most HA subtypes (H1–H15, see methods section for details) with one click that replaces the flexible linker and transmembrane region with a stabilizing Trimerization domain, an Avitag for mono-biotinylation, and a histidine six-mer (H6) sequence for nickel-based purification (Figure 2B)[27]. Using the "probe option" of this function also introduces a "Y98F" mutation (H3 numbering) that reduces binding to sialic acid for probes to be used in cellular assays such as for flow cytometric sorting of HA-specific B cells[28] or Libra-seq[29].

In addition to the mutagenesis function, Librator also provides several sequence editing modes. For example, the Cocktail mode allows users to compare a donor sequence to a template sequence and scan all amino acid differences between them. Then Librator will automatically generate multiple sequences based on the template sequence with each identified mutation or their combinations (Figure 2C). This function improves the efficiency of identifying key residues between antigenically or functionally distinct viruses. Users can instantaneously generate a library of point mutant variants for expression to, for example, identify which amino acids differing between two HA molecules are important for the binding of a monoclonal antibody or drive differential function of the compared HA molecules, such as host-species tropism. With these functions, users can generate demanded mutations in batches in minutes, compared with manual generation of mutated sequences that usually takes at least several hours.

We also implemented a sequence designer in Librator to facilitate complicated sequence design. Current influenza vaccine-antigen design efforts aim to retarget immunity away from some epitopes and focused on others through the production of chimeric and mosaic HA and NA proteins. Compared to individual mutagenesis, chimeric and mosaic sequence design usually requires mutating multiple regions (groups of residues) or even splicing sequences together from multiple influenza strains. Large numbers of mutations and complex design make manual design of chimeric/mosaic HA proteins difficult and prone to error. To overcome this challenge, Librator includes an interactive GUI to enable easy and efficient design of chimeric and mosaic proteins (Figure 2D). Using the graphical sequence viewer, users can easily specify and highlight regions to be replaced on a template sequence and regions to be inserted from a donor sequence. A dedicated viewer displays the current product with information about all replacements. After users review and confirm the current product in the product viewer, Librator will generate a new record of the user-designed product, with nucleotide sequence, subtype (same as template) and mutated residues. Using Librator, biologists can easily design complicated chimeric and mosaic HA/NA sequences with extensive mutations or replacement of entire epitopes or regions. Use of Librator in our lab has enormously improved the efficiency and accuracy of sequence design.

Librator's cloning functions also maximizes the economy, practicality, and accuracy of synthesizing nucleotide sequences for expression by Gibson cloning using a recipe-based generator. For this Librator capitalizes on the fact that Gibson cloning uses sequence homology of a short overlap/joint region (usually 20–25bp) between neighboring fragments and also that most HA and NA sequences of a type have highly conserved and homologous regions interspersed with the variable sequence elements. Natural mutations in these proteins are enriched in only a few highly variable regions (e.g. epitopes, antibody binding sites) (Figure 3A). The cloning algorithm of Librator optimizes fragment design for HA and NA sequences to maximize the reusability of gene fragments. Librator typically produces HA as four fragments (user customizable) or NA as three fragments and databases all previous fragments generated by a lab so that new HA molecules differing in only one fragment can be synthesized by replacing only the single fragment based on an automatically generated recipe specifying the existing fragments in the laboratories inventory and the new sequence to be synthesized (Figure 3B). For example, an escape mutant HA of a particular strain may contain only several amino acid changes within a single antigenic site in one fragment of the construct. If the original variant was designed by Librator and expressed in the lab, the escape variant can now be synthesized at only $1/4^{th}$ the cost. This function become particularly cost-effective when libraries of point-mutants are generated. For this, Librator identifies potential

overlapping regions by locating highly conserved regions based on peptide conservation of all historical HA and NA sequences. These regions are then used to define fragments on a template sequence for each subtype or group of subtypes, ensuring that end compatibility of fragments is unaffected by sporadic mutations, insertions or deletions. In Librator, all query sequences are aligned to the appropriate template sequence to ensure fragments from different batches are subject to the same design, guaranteeing their reusability. Users can clone and express their HA and NA sequences for a reduced cost by reusing fragments in their inventory (Figure 3C). The more sequences users clone, the more comprehensive a fragment inventory they will amass, enabling more fragment reuse and reagent saving. This is extremely beneficial for labs that are investing continuing efforts and resources into influenza research.

According to the evolutionary history of influenza HA subtypes, we designed uniform fragments on the basis of a classic H1 sequence (A/California/7/2009, H1N1) and a classic H3 sequence (A/Aichi/2/1968, H3N2). We aligned all group 1 HAs (H1, H2, H5, H6, H8, H9, H11, H12, H13, H16, H17, and H18) to the H1 template and all group 2 HAs (H3, H4, H7, H10, H14, and H15) to the H3 template for fragment design (Table S1,S2, Figure S2). For NAs, we designed uniform fragments for each subtype by aligning each of the NA sequences to the template of their respective subtypes (Table S1, Figure S3). This template-mapping-based fragment design ensures that all fragments are standardized and not affected by either different batches or sporadic insertion/deletion events (e.g. a deletion Δ130 between pre-1994 and post-1994 human seasonal H1N1, or insertions in the cleavage site of high pathogenic avian H5 and H7)[26, 30, 31]. We applied this system to several applications to validate its effectiveness and compatibility. Lab practices demonstrated that this tool could help to clone and express proteins at a reduced cost. For example, reagent cost was reduced by 54% when expressing proteins with single mutations to investigate the key residues of the antigenic drift between A/HongKong/4801/2014 (H3N2) and A/Switzerland/9715293/2013 (H3N2) influenza viruses (Supplemental Data S1). Even in an extreme case of expressing HAs of 39 representative H3N2 viruses from 1968 to 2018, using Librator design only increased the reagent cost by 4% while generating many reusable gene fragments for future projects (Supplemental Data S2).

With the effectiveness of this method verified by lab practice, we further developed several supporting functions to enable efficient workflow and a smooth user experience. We enabled users to customize the Gibson upstream connector and downstream connector to fit more vectors. Furthermore, we also designed a customizable C-terminal domain/tag region for HA proteins: Trimerization domain + Purification tag (e.g. 6xHisTag) or Trimerization domain+ AviTag + Purification tag (sequences are user customizable) for better end compatibility (Figure 3D). To reduce the risk of error, we designed an interactive GUI on which users can preview their designed fragments before generating all products (Figure 3E). All generated fragments are archived in an SQL-driven database for better data access and management. To facilitate lab reagent stock management, Librator also allows multiple users to connect to a remote MySQL fragment database (Figure 3F). Once fragments are generated by users, Librator searches the current fragment inventory, then generates a list of reusable fragments already in inventory, novel fragments that need to be ordered and recipes for all sequences. An Excel file containing fragment names and sequences in the format of a 96-well plate is also generated and can be sent to a DNA synthesis company directly. FASTA format files that contain the fragments of each sequence are generated as well, enabling users to validate their compatibility using sequence analysis software. Lastly, we

also developed a general fragment design feature that allows users to split any nucleotide sequence into a few customized fragments, most applicable when reusability is not a priority (Figure 3G). This feature will be helpful for novel or frontier research in particular, such as in designing Gibson cloning fragments for novel COVID-19 proteins.

In conclusion, we developed a variety of functions associated with interactive GUIs in Librator, aiming to improve research efficiency and liberate biologists from onerous and repetitive work so that they can focus on more productive aspects of influenza research. This feature greatly facilitates the work of users who are not familiar with command-line tools, as well as reducing the possibility of mistakes. Furthermore, by the help of two widely used structure visualization tools, Librator seamlessly links users linear HA sequences to 3D structures that are annotated by peptide conservation and known epitopes. This unique feature facilitates virologists, especially those who are not expertise in structural biology, to investigate their sequences and designs from structural aspect. We also provide tools for optimized Gibson clone fragment designs for HA and NA proteins of influenza viruses, enabling low-cost protein cloning and expression. This protocol liberates more scientific potentials for related research under limited budgets, expending the depth and breadth of related research. Looking to the future, Librator has much potential to be extended. In recent years, more and more studies have revealed epitopes on NA proteins and highlighted the importance of NA as a target of human antibodies[32-35]. Compared to HA, there is still a lack of knowledge of NA. In the near future, more and more studies will focus on NA and will be able to generate comprehensive profiles of epitopes on NA. Librator will be continuously updated with the latest research progress on NA. Furthermore, compared to influenza A, there is a lack of knowledge about influenza B, which also has an impact on public health and is also an important component of WHO-recommended influenza vaccine formulas. Improving support for influenza B is another future goal for Librator. Lastly, this template-based and standardized fragment design also has the potential to be extended to other viruses, such as human immunodeficiency virus (HIV) or hepatitis C virus (HCV) or coronaviruses. The modularized structure of this software is also ready for secondary development to be compatible with more biological contexts. With this in mind all source code is provided and we encourage updates and feedback and hope that Librator becomes a community-based tool and development effort.

Methods

Dataset

All the HA and NA sequences used in this study were downloaded from the NCBI FLU database (https://www.ncbi.nlm.nih.gov/genomes/FLU/)[36] and GISAID database (https://www.gisaid.org/)[37]. H1 protein: 2243 seasonal H1 sequences and 31575 pdm09 sequences. H3 protein: 61798 sequences. NA protein: 28747 N1 sequences, 15194 N2 sequences, 1430 N3 sequences, 291 N4 sequences, 382 N5 sequences, 2420 N6 sequences, 1188 N7 sequences, 2446 N8 sequences and 2446 N9 sequences. All sequences are peptide sequences.

Gibson Clone fragment design for HA and NA proteins

Gibson Clone fragments should be designed according to a uniform criterion that is unaffected by sporadic insertions/deletions in different strains, and all the joint regions of neighboring fragments should be located at the most conserved region. Furthermore, an optimized fragment design should also balance the reusability of each single fragment and the total number of fragments. The shorter

a single fragment is, the less the probability of mutations will be, enabling higher reusability of each fragment; too short a fragment length will result in a larger number of fragments, however, which highly increases the total reagent cost.

To determine the optimized fragment design (including number of fragments and joint region location), we investigated amino acid variations of all residues of human H1, human H3 and NA (all hosts), and we quantified the amino acid variations by an amino acid variation entropy function[20].

$$S_{obs} = -\sum_{n=1}^{N} p_n log_2 p_n$$

$$p_n = count_n / \sum_{n=1}^{N} count_n$$

S_{obs} denotes the entropy of the observed symbol. p_n denotes the frequency of the n-th amino acid of this residue, N denotes the total number of all possible amino acids (N = 20), and $count_n$ denotes total number of the n-th amino acid of this residue.

By comprehensively considering commercial DNA fragment sizes and prices and distribution of the conserved regions in the HA/NA sequences, we proposed an optimized fragment design that divides HA into 4 fragments and NA into 3 fragments. Length of joint regions was set to 9 amino acids (27bp in nucleotides) because Gibson Clone Assembly requires at least 25bp joint region length. Joint regions of group 1 HA were set at 123–131, 264–272, and 403–411 (CDS position on a A/California/7/2009[H1N1] HA). Joint regions of group 2 HA were set at 123–131, 265–273, and 403–411 (CDS position on a A/Aichi/2/1968[H3N2] HA). Joint regions of NA were set at 131–139 and 292–301 (for each subtype, all positions are subject to CDS position on a representative template of this subtype). Joint regions and templates of all HA and NA subtypes are shown in Table S1. Furthermore, to maximize the compatibility of joint regions, Librator revised all nucleotide sequences of joint regions by translating them from peptide sequences using a dictionary in which each amino acid only has one corresponding codon.

Pipeline design

To optimize the user experience, especially for biologists without a computer science background and not familiar with command-line tools, we developed a highly interactive GUI for Librator. Function calling, parameter setting and information display were integrated into one main interface with multiple tabs. All functions can be divided into two broad categories: basic function and advanced function. Basic function includes Input/output (I/O) operations and database (DB) operations: parameter setting, create new sequence DB, open existing sequence DB, import sequences and export sequences. Advanced function inGUIcludes sequence design/editing, fragment design, phylogenetic analysis and structure visualization: the specific functions are sequence information editing, HA numbering, mutation identification, antigen probe design, multiple sequence alignment, phylogenetic analysis, sequence editing, chimeric HA design, structure visualization and Gibson Clone fragments design (Figure S4).

Multiple HA numbering schemes in Influenza research field

As discussed in the introduction section, there are three different numbering systems commonly used in the Influenza research field: 1) CDS position, 2) crystal-structure-based H1/H3 numbering, and 3) Burke and Smith HA numbering scheme.

Residue number on CDS is usually counted from the first amino acid of the CDS (Methionine). For a given sequence, CDS position can cover all residues of given sequence regardless of sporadic insertion and/or deletion. The crystal-structure-based H1/H3 numbering aligns given sequences against a classic H1/H3 template and assigns position numbers for all residues that can map to the template crystal structures. Thus, inserted residues and non-structural residues (e.g. signal peptides) will not be assigned a residue number because they cannot be aligned to the template crystal structures. Furthermore, numbers of residues in HA1 and HA2 are counted independently. The Burke and Smith HA numbering scheme proposed by Burke et al. aligns given sequences against 26 templates of different subtypes to determine the residue numbers. Different from the structure-based HA numbering scheme, the Burke and Smith HA numbering scheme is based on amino acid sequences without considering structural information, and it counts from the first amino acid of the CDS after signal peptide removal. This numbering scheme has been implemented by FLUDB (https://www.fludb.org/brc/haNumbering.spg?method=ShowCleanInputPage&decorator=influenza) recently. We compared three different HA numbering systems using H1 and H3 template sequences (Figure S5; Table S3, S4).

Because protein structures play an important role in antigen phenotypes, all functions in Librator, including alignment viewer, sequence editing and sequence designing were based on structure-based HA numbering systems. Users can only access the universal HA numbering scheme in the "Burke and Smith HA numbering" viewer.

Antigen probe design

The antigen probe design function makes HA probes for a given HA sequence by generating a "Y98F" mutation (H3 numbering) and replacing the flexible linker and transmembrane region with Trimerization-Avitag-H6 sequence. Residue 98 under H3 numbering is located by the built-in HA numbering aligner system automatically. The transmembrane region is identified by aligning given sequences to an H3 template. This function is not available for most H16 HAs and all H17 and H18 HAs because these sequences are isolated from avian and bat sources, and their residue 98 under H3 numbering is already "F."

Identification of key residues between two groups of sequences

In this function, first we align all sequences from both groups together; then we investigate peptide differences between the two groups for every residue independently. For each residue, we convert amino acid composition of two groups into numerical amino acid vectors:

$$V = \{N_{AA_1}, N_{AA_2}, \dots, N_{AA_{21}}\} / \sum N_{AA_i}$$

AA_i denotes the i-th amino acid of a total of 21 different amino acid options (20 amino acids + any symbol beside those 20 AAs, e.g. alignment gap or unclear amino acid X). N_{AA_i} denotes the total number of appearances of the i-th amino acid. Then we defined a score to represent the difference in amino acid composition between two groups on a specific residue:

$$Score_j = |V_{pos_j} - V_{neg_j}|^2$$

V_{pos_j} and V_{neg_j} denote amino acid vectors of residue j. $Score_j$ denotes peptide difference of residue j.

Under this scoring system, score = 0 indicates no peptide difference on this residue between two groups. The higher the score is, the bigger the peptide difference will be. Then all residues will be ranked by the score from high to low to facilitate users' further analysis. In summary, this function gives suggestions of key residues to narrow the candidate range and accelerate biological studies.

Nomenclature of Gibson Clone fragments

Users can save resources and reagents by reusing standardized fragments generated by Librator. We defined a nomenclature for all fragments for easier inventory management. Each fragment name is composed of three parts: gene segment subtype (H1–H18, N1–N11), fragment number (F1–F4 for HA, F1–F3 for NA) and a unique numerical ID. For example, H3-F1-0001 denotes a gene fragment at position 1 (first fragment) generated from an H3 sequence with an ID 0001. We designed a SQL table for Librator for inventory management of all gene fragments. There are 9 keys in the fragment table: Name (prime key), Segment (HA/NA), Fragment (F1–F4), Subtype, ID, Template (template sequence name), AAseq (amino acid sequence), NTseq (nucleotide sequence), and Instock (yes/no). We also designed an interface for users to manage their fragment inventory.

Implementation

The pipeline was primarily implemented in Python3 (version 3.7.3 for MacOS, version 3.9 for Windows 10) using PyQT5 library (version 5.13.0). The executable application was compiled from source code using Pyinstaller (version 4, https://www.pyinstaller.org/). JQuery JavaScript library (version 3.4.1, https://jquery.com/), pyecharts library (version 1.8.1, https://pyecharts.org/) and matplotlib library (version 3.1.1, https://matplotlib.org/) were used to generate figures and integrative HTML sequence viewers. Local databases were generated by sqlite3 (https://docs.python.org/3/library/sqlite3.html), a Python version of SQLite (version 3.33.0, https://www.sqlite.org/); remote databases were generated by MySQL (version 8.0, https://www.mysql.com/). The entire project was developed using PyCharm CE community (version 2019.2, https://www.jetbrains.com/pycharm/) integrated development environment. We integrated two sequence aligners MUSCLE (version 3.8.31, https://www.drive5.com/muscle/) and Clustal Omega (version 1.2.3, http://www.clustal.org/omega/) for multiple sequence alignment, H1/H3 numbering alignment, fragment alignment and mutation identification [38, 39]. We also implemented an interface for users to visualize their sequences on 3D structures using PyMOL (version 2.3.2, https://pymol.org/) and UCSF Chimera (https://www.cgl.ucsf.edu/chimera/), to generate a maximum-likelihood tree using RAxML (version 8.0.0, https://raxml-ng.vital-it.ch/), and to visualize a phylogenetic tree using phylotree.js library (http://phylotree.hyphy.org/)[21, 22, 40]. Sequence logos of selected sequences were generated by WebLogo (version 3.7.1, https://weblogo.berkeley.edu/) [20]. Landsacpe of multiple sequence alignment is generated by html2canvas (https://html2canvas.hertzen.com/) and Python3. The codon optimization functions is powered by DNA Chisel (version 3.2.6, https://github.com/Edinburgh-Genome-Foundry/DnaChisel). The crystal-structure-based HA numbering system is adopted from a public repository (https://github.com/bloomlab/HA_numbering) with some modifications.

Software and code availability

Librator is freely hosted online (https://wilsonimmunologylab.github.io/Librator/). Tutorials are available from a Wilson Lab GitHub page (https://wilsonimmunologylab.github.io/Librator/), a pdf format user guide is also available for downloading. The source code is also available from GitHub (https://github.com/WilsonImmunologyLab/Librator).

We provide executable version of this software for two dominated operating systems: Windows 10 and MacOS. The MacOS version of this software is compiled under macOS Mojave (version 10.14.6) and has been tested under macOS Mojave (version 10.14.6), macOS Catalina (version 10.15.2) and macOS Big Sur (version 11.2.3). The Windows 10 version of this software is compiled under Windows 10 Home (OS build 19042.867) and has been tested under the same system.

This Python-based software is also transferable and can be compiled under other systems (e.g. ubuntu) from source code.

Acknowledgements

We would like to thank Dr. Jesse Bloom for his assistance and suggestions for this project.

FUNDING

This project was funded in part by the National Institute of Allergy and Infectious Disease (NIAID); National Institutes of Health (NIH) grant numbers U19AI082724 (P.C.W.), U19AI109946 (P.C.W.), U19AI057266 (P.C.W.), and the NIAID Centers of Excellence for Influenza Research and Surveillance (CEIRS) grant numbers HHSN272201400005C (P.C.W.).

Author Contributions

L.L. designed the model, implemented the software, performed computational analyses, and wrote the manuscript. O.S., J.J.G., S.C. and Y.F. tested software, improved software design, and revised the manuscript. H.L.D. and C.T.S. tested software and improved software design. N.Z. and M.H. performed experimental validations. P.C.W. initiated and supervised the work, designed the model, implemented the software, and wrote the manuscript.

Competing Interests

The authors declare no competing interests.

References

1. Koel, B.F. et al. Substitutions near the receptor binding site determine major antigenic change during influenza virus evolution. *Science* **342**, 976-979 (2013).
2. Sun, W. et al. Development of influenza B universal vaccine candidates using the "Mosaic" hemagglutinin approach. *Journal of virology* **93** (2019).

3. Krammer, F. & Palese, P. Universal influenza virus vaccines that target the conserved hemagglutinin stalk and conserved sites in the head domain. *The Journal of infectious diseases* **219**, S62-S67 (2019).
4. Carter, D.M. et al. Design and characterization of a computationally optimized broadly reactive hemagglutinin vaccine for H1N1 influenza viruses. *Journal of virology* **90**, 4720-4734 (2016).
5. Gibson, D.G. et al. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nature methods* **6**, 343-345 (2009).
6. Burney, L.E. Influenza immunization: statement. *Public health reports* **75**, 944 (1960).
7. Flannery, B. et al. Interim estimates of 2017–18 seasonal influenza vaccine effectiveness—United States, February 2018. *Morbidity and Mortality Weekly Report* **67**, 180 (2018).
8. Xie, H. et al. H3N2 mismatch of 2014–15 northern hemisphere influenza vaccines and head-to-head comparison between human and ferret antisera derived antigenic maps. *Scientific reports* **5**, 1-10 (2015).
9. Chiu, C. et al. Cross-reactive humoral responses to influenza and their implications for a universal vaccine. *Annals of the New York Academy of Sciences* **1283**, 13-21 (2013).
10. Hagan, T. et al. Antibiotics-driven gut microbiome perturbation alters immunity to vaccines in humans. *Cell* **178**, 1313-1328. e1313 (2019).
11. Henry, C. et al. Monoclonal Antibody Responses after Recombinant Hemagglutinin Vaccine versus Subunit Inactivated Influenza Virus Vaccine: a Comparative Study. *Journal of Virology* **93**, e01150-01119 (2019).
12. Staneková, Z. & Varečková, E. Conserved epitopes of influenza A virus inducing protective immunity and their prospects for universal vaccine development. *Virology journal* **7**, 1-13 (2010).
13. Burke, D.F. & Smith, D.J. A recommended numbering scheme for influenza A HA subtypes. *PloS one* **9** (2014).
14. Kirkpatrick, E., Qiu, X., Wilson, P.C., Bahl, J. & Krammer, F. The influenza virus hemagglutinin head evolves faster than the stalk domain. *Scientific reports* **8**, 1-14 (2018).
15. Deem, M.W. & Pan, K. The epitope regions of H1-subtype influenza A, with application to vaccine efficacy. *Protein Engineering, Design & Selection* **22**, 543-546 (2009).
16. Hai, R. et al. Influenza viruses expressing chimeric hemagglutinins: globular head and stalk domains derived from different subtypes. *Journal of virology* **86**, 5774-5781 (2012).
17. Steel, J. et al. Influenza virus vaccine based on the conserved hemagglutinin stalk domain. *MBio* **1** (2010).
18. Medina, R.A. et al. Glycosylations in the globular head of the hemagglutinin protein modulate the virulence and antigenic properties of the H1N1 influenza viruses. *Science translational medicine* **5**, 187ra170-187ra170 (2013).
19. Li, L. et al. Multi-task learning sparse group lasso: a method for quantifying antigenicity of influenza A (H1N1) virus using mutations and variations in glycosylation of Hemagglutinin. *BMC bioinformatics* **21**, 1-22 (2020).
20. Crooks, G.E., Hon, G., Chandonia, J.-M. & Brenner, S.E. WebLogo: a sequence logo generator. *Genome research* **14**, 1188-1190 (2004).
21. Schrödinger, L. The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC (2017). *Google Scholar There is no corresponding record for this reference*.

22. Pettersen, E.F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *Journal of computational chemistry* **25**, 1605-1612 (2004).
23. Abola, E.E., Bernstein, F.C. & Koetzle, T.F. in Neutrons in Biology 441-441 (Springer, 1984).
24. Weis, W.I., Brunger, A.T., Skehel, J.J. & Wiley, D.C. Refinement of the influenza virus hemagglutinin by simulated annealing. *J Mol Biol* **212**, 737-761 (1990).
25. Zhang, W. et al. Molecular basis of the receptor binding specificity switch of the hemagglutinins from both the 1918 and 2009 pandemic influenza A viruses by a D225G substitution. *J Virol* **87**, 5949-5958 (2013).
26. McDonald, N.J., Smith, C.B. & Cox, N.J. Antigenic drift in the evolution of H1N1 influenza A viruses resulting from deletion of a single amino acid in the haemagglutinin gene. *Journal of General Virology* **88**, 3209-3213 (2007).
27. Benton, D.J. et al. Influenza hemagglutinin membrane anchor. *Proceedings of the National Academy of Sciences* **115**, 10112-10117 (2018).
28. Whittle, J.R. et al. Flow cytometry reveals that H5N1 vaccination elicits cross-reactive stem-directed antibodies from multiple Ig heavy-chain lineages. *J Virol* **88**, 4047-4057 (2014).
29. Setliff, I. et al. High-Throughput Mapping of B Cell Receptor Sequences to Antigen Specificity. *Cell* **179**, 1636-1646 e1615 (2019).
30. Harvey, W.T. et al. Identification of low-and high-impact hemagglutinin amino acid substitutions that drive antigenic drift of influenza A (H1N1) viruses. *PLoS pathogens* **12** (2016).
31. Li, L., DeLiberto, T.J., Killian, M.L., Torchetti, M.K. & Wan, X.-F. Evolutionary pathway for the 2017 emergence of a novel highly pathogenic avian influenza A (H7N9) virus among domestic poultry in Tennessee, United States. *Virology* **525**, 32-39 (2018).
32. Krammer, F., Li, L. & Wilson, P.C. Emerging from the Shadow of Hemagglutinin: Neuraminidase Is an Important Target for Influenza Vaccination. *Cell Host & Microbe* **26**, 712-713 (2019).
33. Zhu, X. et al. Structural basis of protection against H7N9 influenza virus by human anti-N9 neuraminidase antibodies. *Cell host & microbe* **26**, 729-738. e724 (2019).
34. Gilchuk, I.M. et al. Influenza H7N9 virus neuraminidase-specific human monoclonal antibodies inhibit viral egress and protect from lethal influenza infection in mice. *Cell host & microbe* **26**, 715-728. e718 (2019).
35. Chen, Y.-Q. et al. Influenza infection in humans induces broadly cross-reactive and protective neuraminidase-reactive antibodies. *Cell* **173**, 417-429. e410 (2018).
36. Bao, Y. et al. The influenza virus resource at the National Center for Biotechnology Information. *Journal of virology* **82**, 596-601 (2008).
37. Bogner, P., Capua, I., Lipman, D.J. & Cox, N.J. A global initiative on sharing avian flu data. *Nature* **442**, 981-981 (2006).
38. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* **32**, 1792-1797 (2004).
39. Sievers, F. et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular systems biology* **7** (2011).
40. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-1313 (2014).

FIGURES

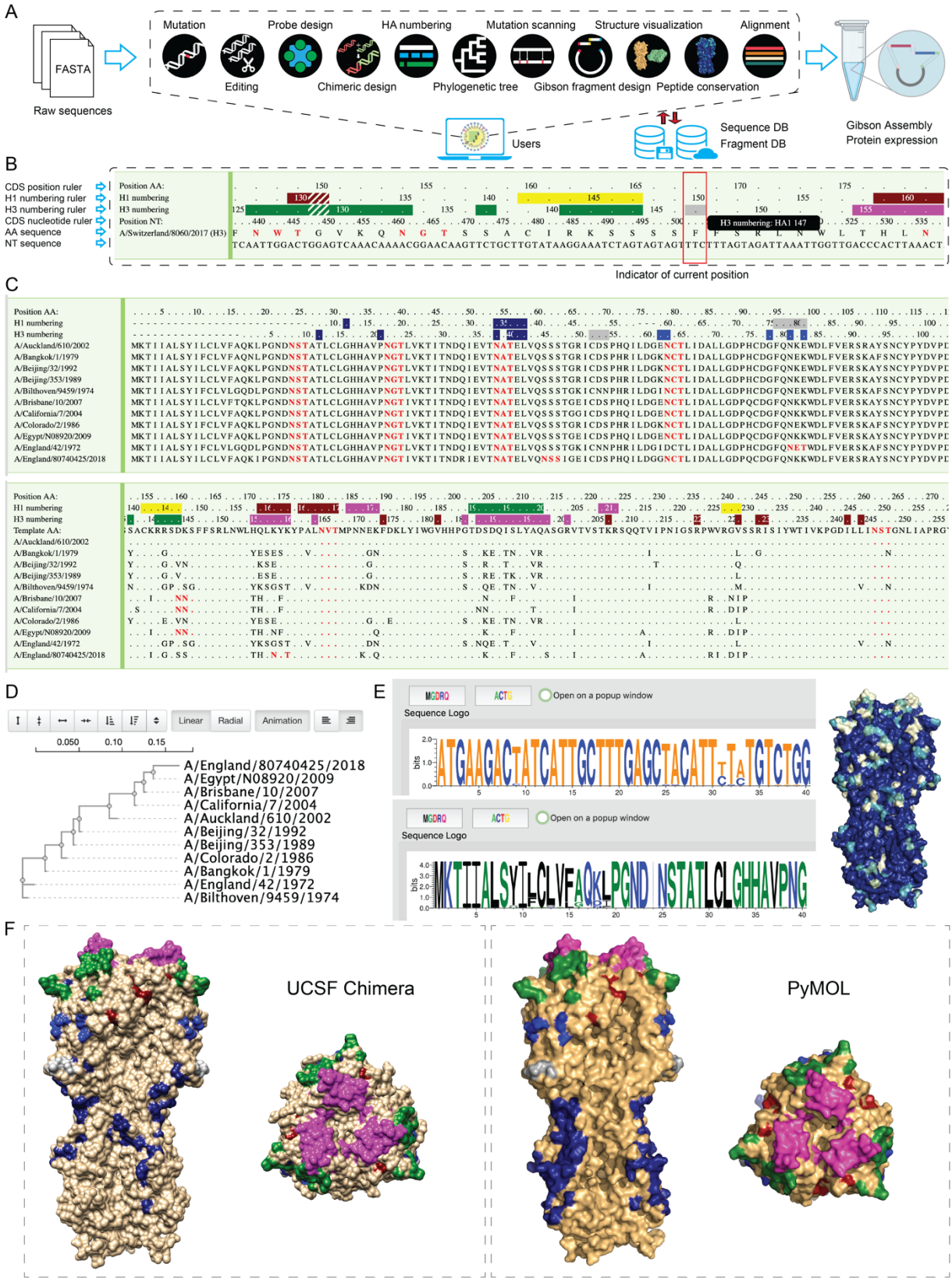

Figure 1. Librator enables efficient analysis of HA and NA influenza virus protein sequences with a variety of functions and a graphical user interface. (**A**) Librator seamlessly connects nucleotide sequences from public databases and lab work, providing a variety of functions for sequence editing and design. (**B**) Librator's HA numbering aligner is integrated in a graphical viewer. Three numbering rulers—a CDS position ruler, H1 numbering ruler and H3 numbering ruler—indicate position information for selected residues. (**C**) Multiple sequence alignment viewer. Original sequence mode is displayed on the top, and template mode is displayed on the bottom. (**D**) Phylogenetic analysis function and tree viewer. Users are allowed to generate phylogenetic trees using either nucleotide sequences or peptide sequences. (**E**) Librator allows users to assess nucleotide conservation and peptide conservation. Librator also can visualize the peptide conservation on HA 3D structures with the help of PyMOL and USCF Chimera. (**F**) Librator allows users to visualize peptides on 3D structures of HA proteins with color-annotated amino acids at all antigenic regions and user-defined sequence labels with the help of PyMOL and USCF Chimera.

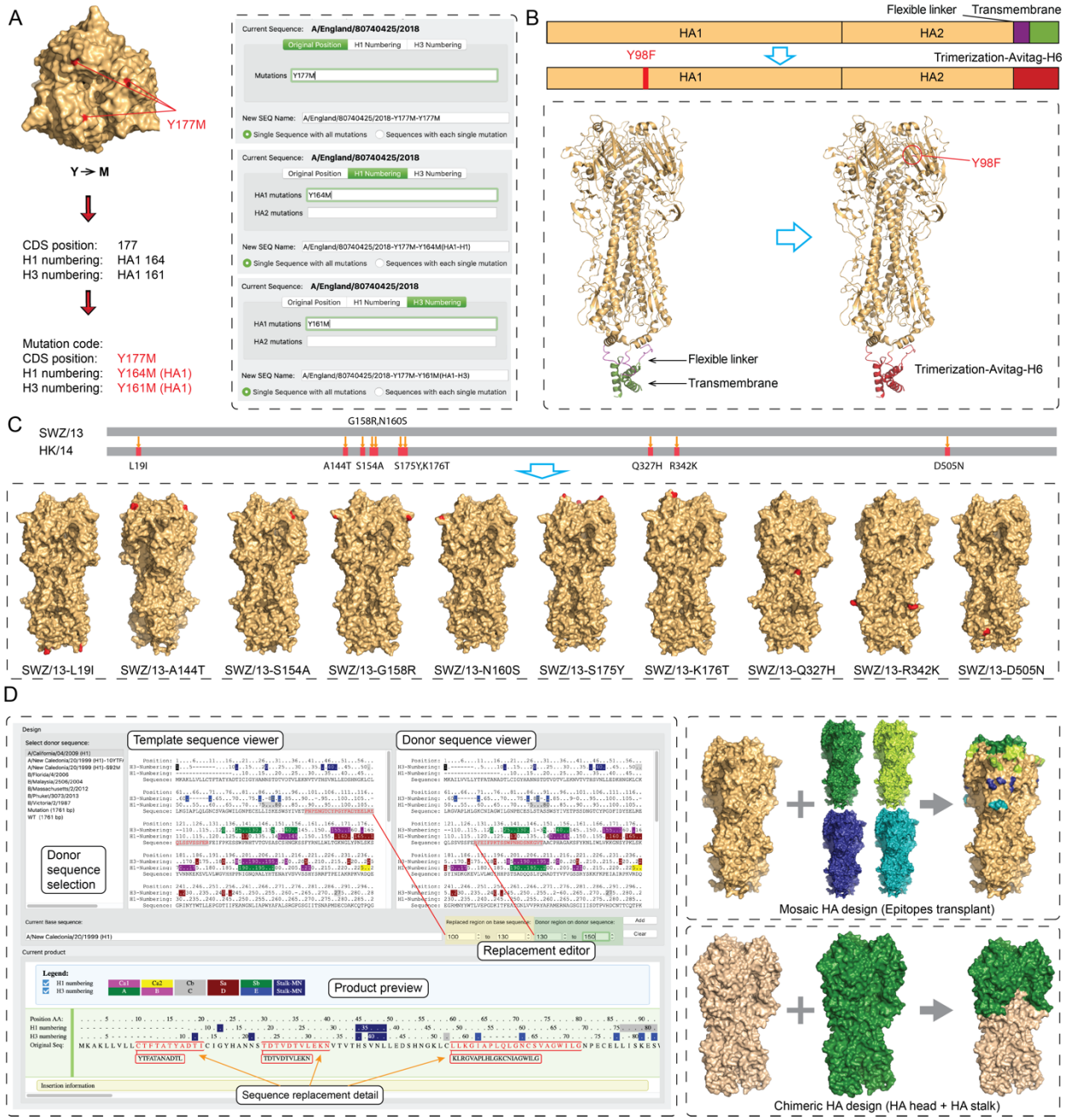

Figure 2. Librator enables efficient design of HA and NA influenza virus proteins. (**A**) Demonstration of mutating a residue on an HA sequence using three numbering systems in Librator. (**B**) Making an antigen probe for a given HA sequence. Librator designs antigen probes for given HA sequences by generating the mutation Y98F (H3 numbering) and replacing the flexible linker and transmembrane region with a Trimerization-Avitag-H6 sequence. This process is demonstrated using an HA structure of A/duck/Alberta/35/76 (H1N1, PDB ID: 6HJR). (**C**) Scanning all amino acid differences between two antigenically distinct sequences (A/Switzerland/9715293/2013 [SWZ/13] and A/HongKong/4801/2014 [HK/14]) with Librator generates a series of sequences, each with a single mutation, to identify key residues of the antigenic drift. (**D**) Designing chimeric sequences using Librator. Users can replace regions on the target sequence with regions from multiple donor sequences. Details of the product can be previewed on a graphical viewer. This function is designed to transplant epitopes from one sequence (or multiple sequences) to another or to combine the HA1 (HA head) from one sequence and HA2 (HA stalk) from another.

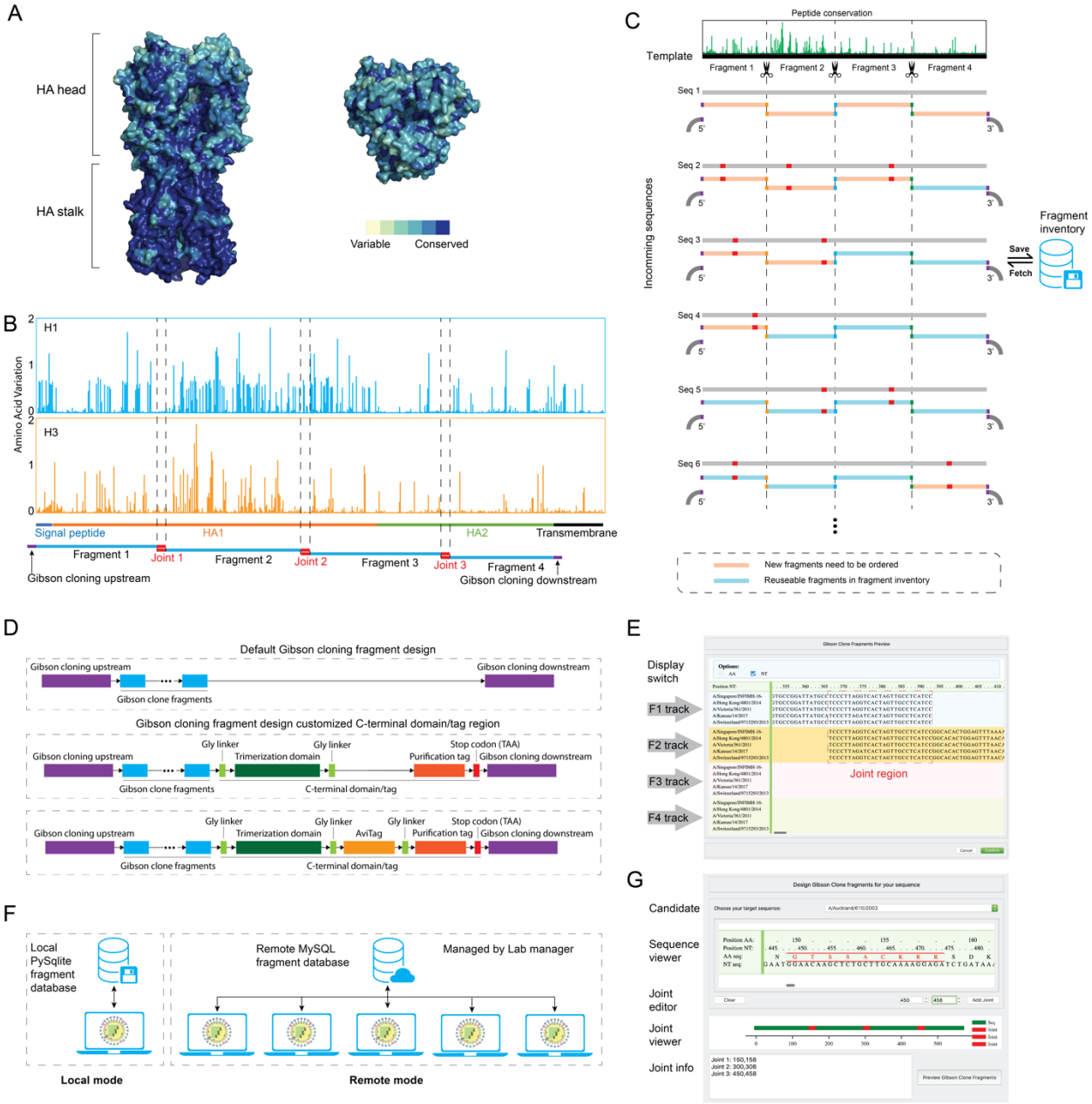

Figure 3. Librator helps users to save reagent cost by designing optimized Gibson Clone fragments for HA and NA sequences. (**A**) Natural mutations on HA protein are enriched in a few highly variable regions. Peptide conservation was visualized on a H1 protein structure (A/California/04/2009 H1N1, PDB ID: 4JTV). Peptide conservation was calculated from HAs of 58 representative H1N1 viruses from 1918 – 2018. (**B**) Illustration of fragment designs for a group 1 HA (based on a H1 template) and a group 2 HA (based on a H3 template) in Librator. Joint regions were determined by locating highly conserved regions on H1/H3 peptides and balancing the length of each fragment. (**C**) Librator determines overlapping regions based on peptide conservation of all historical HA and NA sequences, and then defines fragment design on a template sequence for each subtype. For each given sequence, Librator aligns it to its template sequence to maximize the reusability and compatibility of gene fragments. All fragments are saved in a fragment inventory for further inquiry. Librator aims to save reagent cost by reusing gene fragments. (**D**) Three modes of customizing the C-terminal domain/tag region for the Gibson cloning downstream end. Beside the default mode that directly links the last fragment and the Gibson cloning downstream sequence, we also designed a customizable C-terminal domain/tag region for HA proteins: Trimerization domain + Purification tag (e.g. 6xHisTag) or Trimerization domain + AviTag + Purification tag. (**E**) Graphical viewer of fragments for users to preview their products. (**F**) Librator users can communicate with a local fragment database or a remote fragment database managed

by their lab manager. The remote mode enables better data access and lab reagent stock management. (**G**) Customized fragment design function for any given sequence. This function allows users to add at most 12 joint regions in their sequences and split their sequences into a few fragments for Gibson Assembly. This function was designed for non-influenza sequences or novel research in which reusability is not a priority.

SUPPLEMENTAL FIGURES

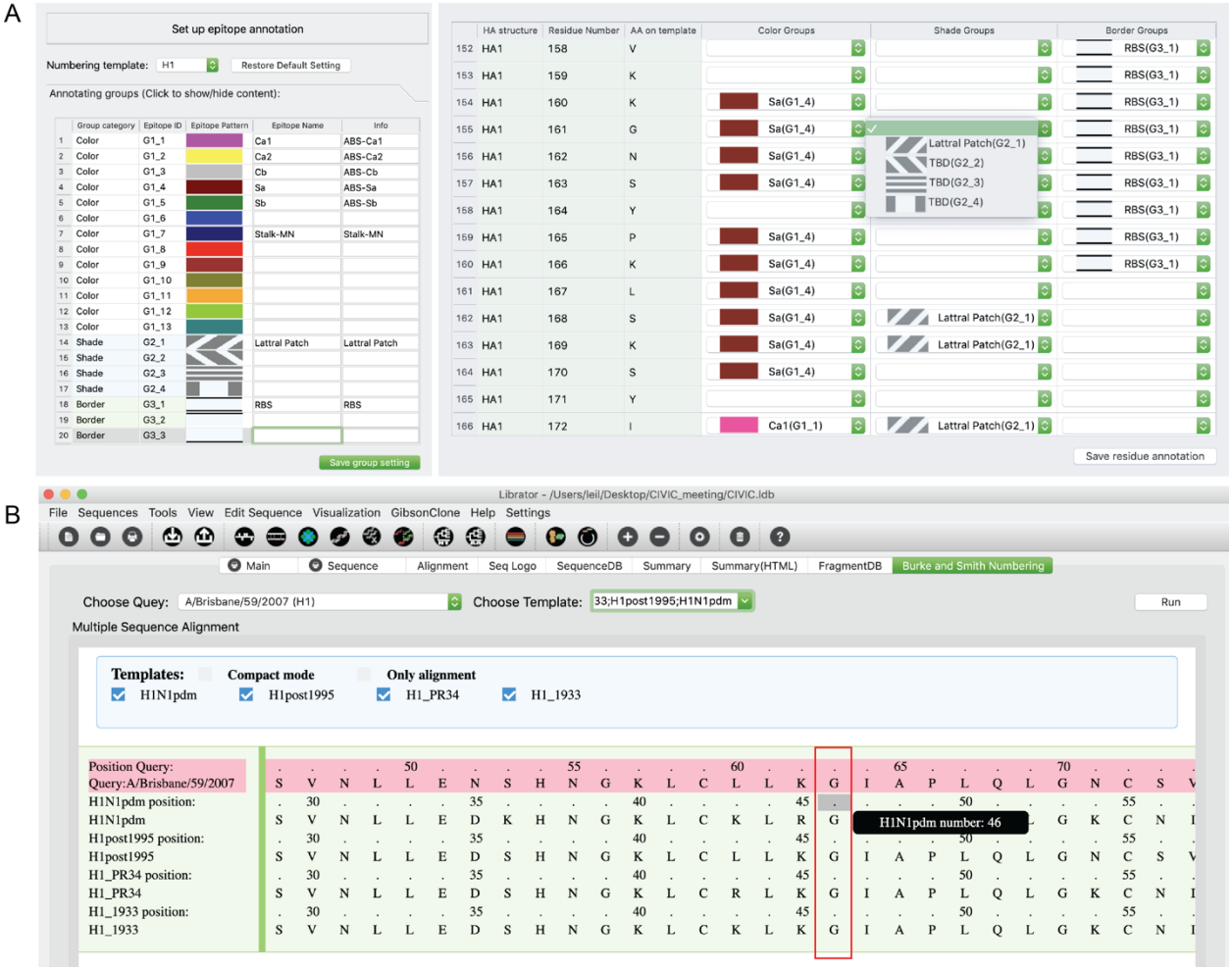

Figure S1. Built-in functions of Librator for influenza sequence analysis. (**A**) User customizable epitopes definition in Librator. Librator allows users to use 13 distinct colors, 4 shade patterns and 3 border styles to annotate individual residues on the sequence viewer according to their specific research interest and focus. The left panel showed GUI of defining epitopes in Librator, and the right panel showed GUI of annotating individual residues using user-defined epitopes. (**B**) Burke and Smith HA numbering scheme viewer in Librator. Users are allowed to align query sequence against multiple templates to access residue numbers on different templates.

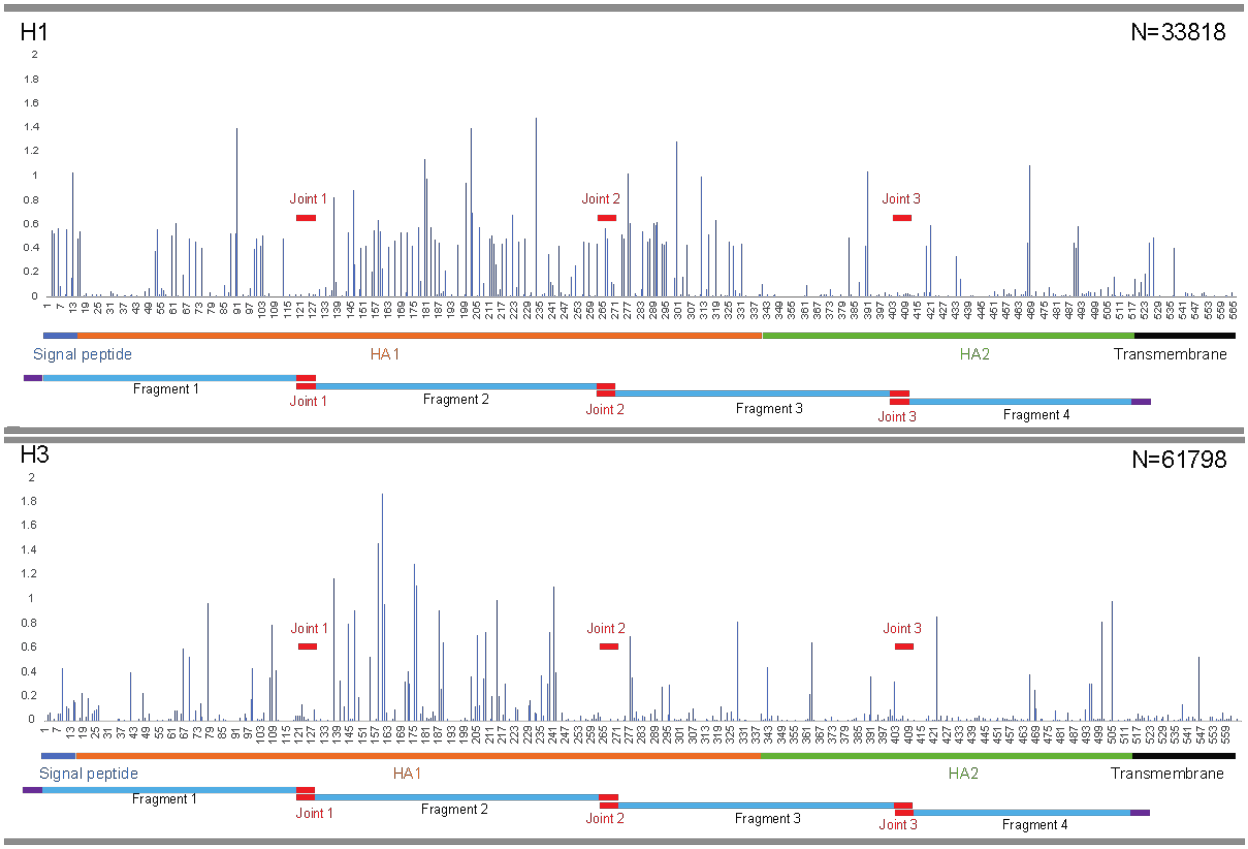

Figure S2. Fragment design for influenza HA proteins. HA proteins are clustered into two groups: group 1 and group 2. In Librator, all group 1 sequences are aligned to an H1 template, and all group 2 sequences are aligned to an H3 template.

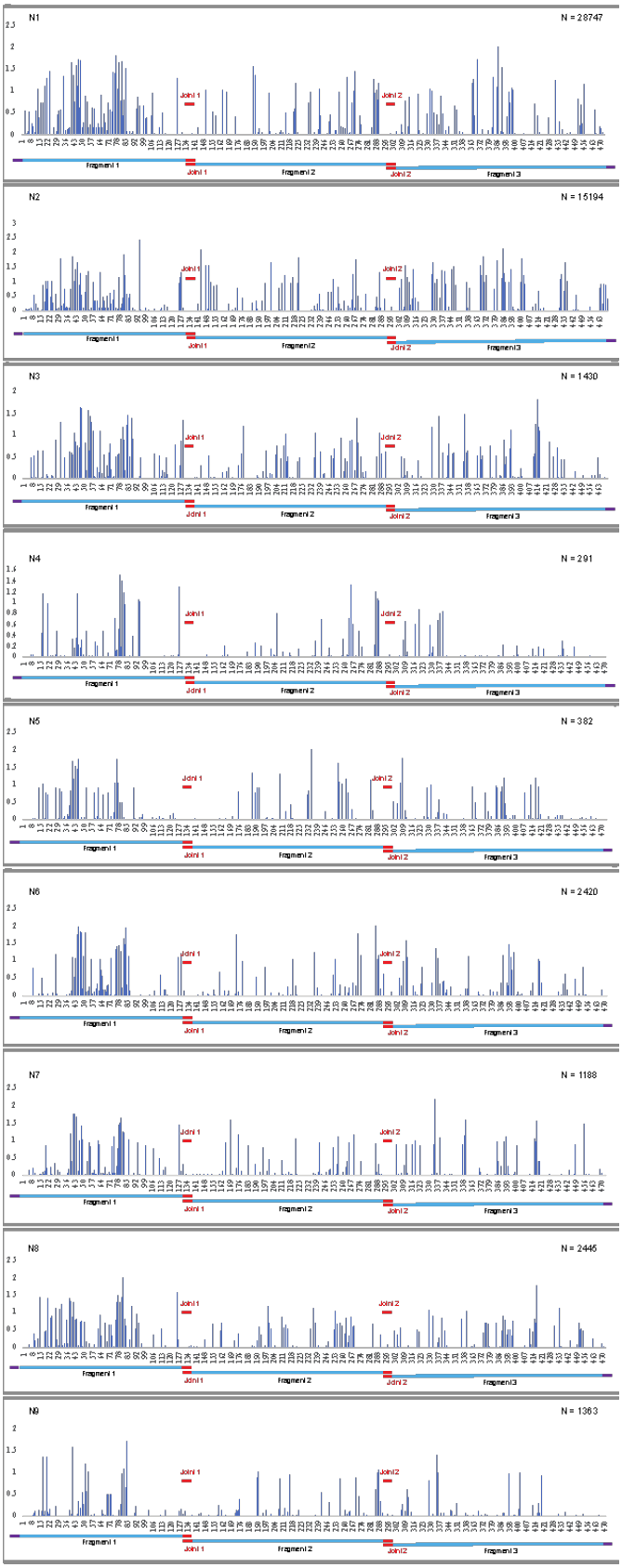

Figure S3. Fragment design for influenza NA proteins.

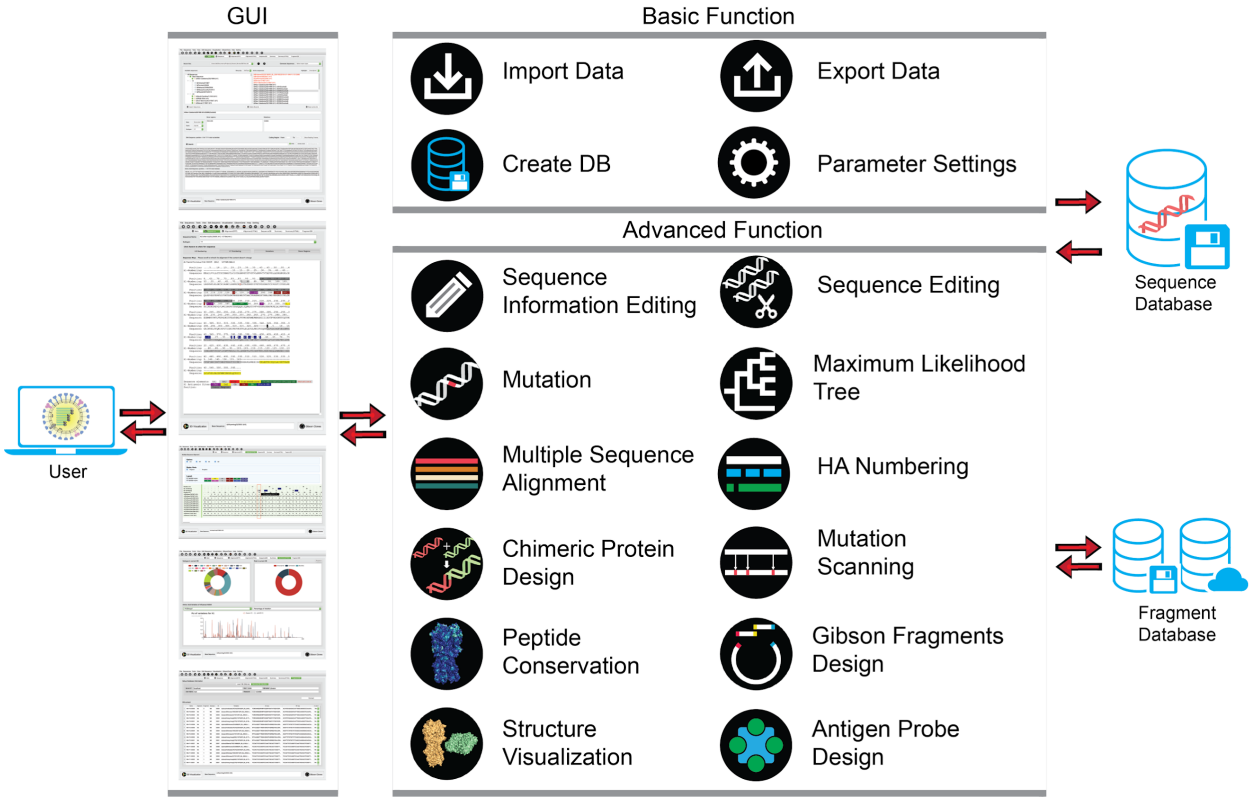

Figure S4. System structure and functions of Librator. Librator is comprised of a UI layer (GUI), logical layer (all functions) and data layer (SQL databases). Users are allowed to finish all operations using the GUI. All functions can be divided into two broad categories: basic function and advanced function. Basic function includes I/O operations and database (DB) operations, and advanced function includes sequence design/editing, fragment design, phylogenetic analysis and structure visualization.

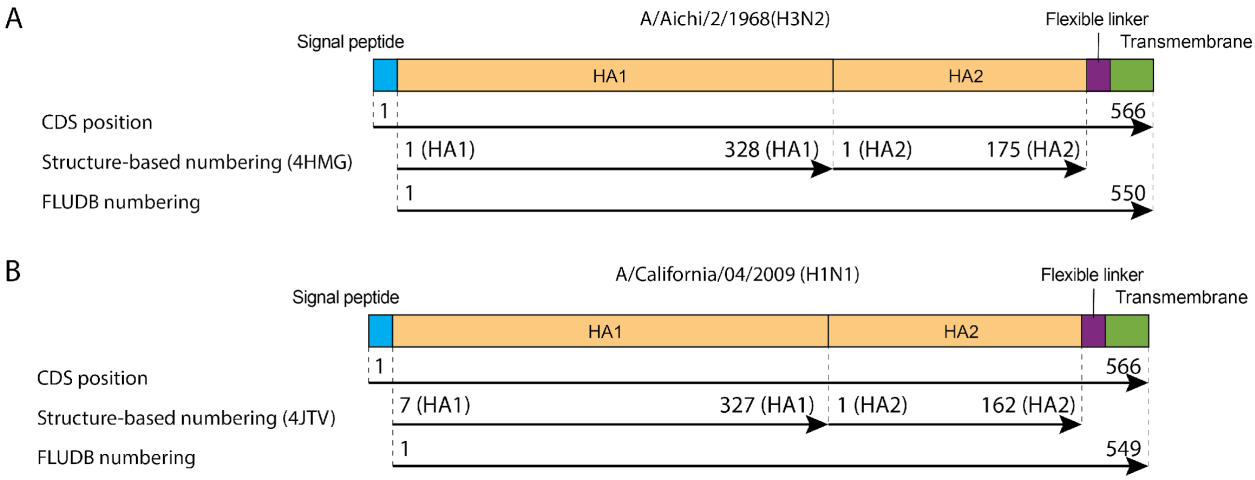

Figure S5. Comparison of three different HA numbering systems using a classic H1 (A/California/04/2009, H1N1) and a classic H3 (A/Aichi/2/1968, H3N2) sequence.

SUPPLEMENTAL TABLES

Table S1. Template sequences and joint region design of HA and NA sequences in Librator.

Protein	Subtype	Template	Joint 1 ***	Joint 2 ***	Joint 3 ***
HA	Group 1*	A/California/04/2009|H1N1	123-131	264-272	403-411
	Group 2**	A/Aichi/2/1968|H3N2	123-131	265-273	403-411
NA	N1	A/California/04/2009|QEH91764	131-139	292-300	
	N2	A/Texas/08/2019|QBP39026	131-139	292-300	
	N3	A/mallard/Maryland/13OS3019/2014|AKF18550	131-139	292-300	
	N4	A/mallard/Utah/AH0020452/2015|AQS26331	131-139	292-300	
	N5	A/commonredshank/Singapore/F83-1/2015|ALR83194	131-139	292-300	
	N6	A/duck/Guangdong/G1345/2014|AJS16549	131-139	292-300	
	N7	A/mallard/Sweden/124987/2010|AHZ37263	131-139	292-300	
	N8	A/northernpintail/Alaska/UGAI15-7291/2015|AOX49352	131-139	292-300	
	N9	A/green-winged-teal/Ohio/14OS1103/2014|AMQ30738	131-139	292-300	
	N10	A/little-yellow-shouldered-bat/Guatemala/060/2010|EPI_ISL_105896	131-139	292-300	
	N11	A/flat-faced-bat/Peru/033/2010|AGX84936	131-139	292-300	

* Group 1 HAs include H1, H2, H5, H6, H8, H9, H11, H12, H13, H16, H17, and H18
** Group 2 HAs include H3, H4, H7, H10, H14, and H15
*** Numbering counts from first amino acid (M) of CDS of H1/H3 template sequence

Table S2. Fragment design for Group 1 HA and Group 2 HA protein sequences.

Subtype	Fragment	Start*	End*	Length (AA)	Length (NT)	Antigenic sites
Group 1 (H1)	Fragment 1	1	131	131	393	Cb, Sb, Stalk-MN
	Fragment 2	123	272	150	450	Ca1, Ca2, Sa
	Fragment 3	264	411	148	444	Cb, Stalk-MN
	Fragment 4	403	518	116	348	Stalk-MN
Group 2 (H3)	Fragment 1	1	131	131	393	C, E, Stalk-MN
	Fragment 2	123	273	151	453	A, B, D
	Fragment 3	265	411	147	441	C, Stalk-MN
	Fragment 4	403	520	118	354	Stalk-MN

* Numbering counts from first amino acid (M) of CDS of H1/H3 template sequence

Table S3. Comparison of multiple HA numbering schemes using pdm09 H1 template (A/California/04/2009, H1N1). Template for FLU DB numbering is H1N1pdm. Flexible linker and transmembrane domain were located by aligning to A/duck/Alberta/35/76(H1N1), PDB ID: 6HJR.

position on CDS	FLU DB H1pdm numbering	structure-based H1 numbering (4JTV)	Residue	Annotation
1	-	-	M	Signal peptide
2	-	-	K	Signal peptide
3	-	-	A	Signal peptide
4	-	-	I	Signal peptide
5	-	-	L	Signal peptide

6	-		-	V	Signal peptide
7	-		-	V	Signal peptide
8	-		-	L	Signal peptide
9	-		-	L	Signal peptide
10	-		-	Y	Signal peptide
11	-		-	T	Signal peptide
12	-		-	F	Signal peptide
13	-		-	A	Signal peptide
14	-		-	T	Signal peptide
15	-		-	A	Signal peptide
16	-		-	N	Signal peptide
17	-		-	A	Signal peptide
18		1	7 (HA1)	D	HA1
19		2	8 (HA1)	T	HA1
20		3	9 (HA1)	L	HA1
21		4	10 (HA1)	C	HA1
22		5	11 (HA1)	I	HA1
23		6	12 (HA1)	G	HA1
24		7	13 (HA1)	Y	HA1
25		8	14 (HA1)	H	HA1
26		9	15 (HA1)	A	HA1
27		10	16 (HA1)	N	HA1
28		11	17 (HA1)	N	HA1
29		12	18 (HA1)	S	HA1
30		13	19 (HA1)	T	HA1
31		14	20 (HA1)	D	HA1
32		15	21 (HA1)	T	HA1
33		16	22 (HA1)	V	HA1
34		17	23 (HA1)	D	HA1
35		18	24 (HA1)	T	HA1
36		19	25 (HA1)	V	HA1
37		20	26 (HA1)	L	HA1
38		21	27 (HA1)	E	HA1
39		22	28 (HA1)	K	HA1
40		23	29 (HA1)	N	HA1
41		24	30 (HA1)	V	HA1
42		25	31 (HA1)	T	HA1
43		26	32 (HA1)	V	HA1
44		27	33 (HA1)	T	HA1
45		28	34 (HA1)	H	HA1

46	29	35 (HA1)	S	HA1
47	30	36 (HA1)	V	HA1
48	31	37 (HA1)	N	HA1
49	32	38 (HA1)	L	HA1
50	33	39 (HA1)	L	HA1
51	34	40 (HA1)	E	HA1
52	35	41 (HA1)	D	HA1
53	36	42 (HA1)	K	HA1
54	37	43 (HA1)	H	HA1
55	38	44 (HA1)	N	HA1
56	39	45 (HA1)	G	HA1
57	40	46 (HA1)	K	HA1
58	41	47 (HA1)	L	HA1
59	42	48 (HA1)	C	HA1
60	43	49 (HA1)	K	HA1
61	44	50 (HA1)	L	HA1
62	45	51 (HA1)	R	HA1
63	46	52 (HA1)	G	HA1
64	47	53 (HA1)	V	HA1
65	48	54 (HA1)	A	HA1
66	49	55 (HA1)	P	HA1
67	50	56 (HA1)	L	HA1
68	51	57 (HA1)	H	HA1
69	52	58 (HA1)	L	HA1
70	53	59 (HA1)	G	HA1
71	54	60 (HA1)	K	HA1
72	55	61 (HA1)	C	HA1
73	56	62 (HA1)	N	HA1
74	57	63 (HA1)	I	HA1
75	58	64 (HA1)	A	HA1
76	59	65 (HA1)	G	HA1
77	60	66 (HA1)	W	HA1
78	61	67 (HA1)	I	HA1
79	62	68 (HA1)	L	HA1
80	63	69 (HA1)	G	HA1
81	64	70 (HA1)	N	HA1
82	65	71 (HA1)	P	HA1
83	66	72 (HA1)	E	HA1
84	67	73 (HA1)	C	HA1
85	68	74 (HA1)	E	HA1

86	69	75 (HA1)	S	HA1
87	70	76 (HA1)	L	HA1
88	71	77 (HA1)	S	HA1
89	72	78 (HA1)	T	HA1
90	73	79 (HA1)	A	HA1
91	74	80 (HA1)	S	HA1
92	75	81 (HA1)	S	HA1
93	76	82 (HA1)	W	HA1
94	77	83 (HA1)	S	HA1
95	78	84 (HA1)	Y	HA1
96	79	85 (HA1)	I	HA1
97	80	86 (HA1)	V	HA1
98	81	87 (HA1)	E	HA1
99	82	88 (HA1)	T	HA1
100	83	89 (HA1)	P	HA1
101	84	90 (HA1)	S	HA1
102	85	91 (HA1)	S	HA1
103	86	92 (HA1)	D	HA1
104	87	93 (HA1)	N	HA1
105	88	94 (HA1)	G	HA1
106	89	95 (HA1)	T	HA1
107	90	96 (HA1)	C	HA1
108	91	97 (HA1)	Y	HA1
109	92	98 (HA1)	P	HA1
110	93	99 (HA1)	G	HA1
111	94	100 (HA1)	D	HA1
112	95	101 (HA1)	F	HA1
113	96	102 (HA1)	I	HA1
114	97	103 (HA1)	D	HA1
115	98	104 (HA1)	Y	HA1
116	99	105 (HA1)	E	HA1
117	100	106 (HA1)	E	HA1
118	101	107 (HA1)	L	HA1
119	102	108 (HA1)	R	HA1
120	103	109 (HA1)	E	HA1
121	104	110 (HA1)	Q	HA1
122	105	111 (HA1)	L	HA1
123	106	112 (HA1)	S	HA1
124	107	113 (HA1)	S	HA1
125	108	114 (HA1)	V	HA1

126	109	115 (HA1)	S	HA1
127	110	116 (HA1)	S	HA1
128	111	117 (HA1)	F	HA1
129	112	118 (HA1)	E	HA1
130	113	119 (HA1)	R	HA1
131	114	120 (HA1)	F	HA1
132	115	121 (HA1)	E	HA1
133	116	122 (HA1)	I	HA1
134	117	123 (HA1)	F	HA1
135	118	124 (HA1)	P	HA1
136	119	125 (HA1)	K	HA1
137	120	126 (HA1)	T	HA1
138	121	127 (HA1)	S	HA1
139	122	128 (HA1)	S	HA1
140	123	129 (HA1)	W	HA1
141	124	130 (HA1)	P	HA1
142	125	131 (HA1)	N	HA1
143	126	132 (HA1)	H	HA1
144	127	133 (HA1)	D	HA1
145	128	134 (HA1)	S	HA1
146	129	135 (HA1)	N	HA1
147	130	136 (HA1)	K	HA1
148	131	137 (HA1)	G	HA1
149	132	138 (HA1)	V	HA1
150	133	139 (HA1)	T	HA1
151	134	140 (HA1)	A	HA1
152	135	141 (HA1)	A	HA1
153	136	142 (HA1)	C	HA1
154	137	143 (HA1)	P	HA1
155	138	144 (HA1)	H	HA1
156	139	145 (HA1)	A	HA1
157	140	146 (HA1)	G	HA1
158	141	147 (HA1)	A	HA1
159	142	148 (HA1)	K	HA1
160	143	149 (HA1)	S	HA1
161	144	150 (HA1)	F	HA1
162	145	151 (HA1)	Y	HA1
163	146	152 (HA1)	K	HA1
164	147	153 (HA1)	N	HA1
165	148	154 (HA1)	L	HA1

166	149	155 (HA1)	I	HA1
167	150	156 (HA1)	W	HA1
168	151	157 (HA1)	L	HA1
169	152	158 (HA1)	V	HA1
170	153	159 (HA1)	K	HA1
171	154	160 (HA1)	K	HA1
172	155	161 (HA1)	G	HA1
173	156	162 (HA1)	N	HA1
174	157	163 (HA1)	S	HA1
175	158	164 (HA1)	Y	HA1
176	159	165 (HA1)	P	HA1
177	160	166 (HA1)	K	HA1
178	161	167 (HA1)	L	HA1
179	162	168 (HA1)	S	HA1
180	163	169 (HA1)	K	HA1
181	164	170 (HA1)	S	HA1
182	165	171 (HA1)	Y	HA1
183	166	172 (HA1)	I	HA1
184	167	173 (HA1)	N	HA1
185	168	174 (HA1)	D	HA1
186	169	175 (HA1)	K	HA1
187	170	176 (HA1)	G	HA1
188	171	177 (HA1)	K	HA1
189	172	178 (HA1)	E	HA1
190	173	179 (HA1)	V	HA1
191	174	180 (HA1)	L	HA1
192	175	181 (HA1)	V	HA1
193	176	182 (HA1)	L	HA1
194	177	183 (HA1)	W	HA1
195	178	184 (HA1)	G	HA1
196	179	185 (HA1)	I	HA1
197	180	186 (HA1)	H	HA1
198	181	187 (HA1)	H	HA1
199	182	188 (HA1)	P	HA1
200	183	189 (HA1)	S	HA1
201	184	190 (HA1)	T	HA1
202	185	191 (HA1)	S	HA1
203	186	192 (HA1)	A	HA1
204	187	193 (HA1)	D	HA1
205	188	194 (HA1)	Q	HA1

206	189	195 (HA1)	Q	HA1
207	190	196 (HA1)	S	HA1
208	191	197 (HA1)	L	HA1
209	192	198 (HA1)	Y	HA1
210	193	199 (HA1)	Q	HA1
211	194	200 (HA1)	N	HA1
212	195	201 (HA1)	A	HA1
213	196	202 (HA1)	D	HA1
214	197	203 (HA1)	T	HA1
215	198	204 (HA1)	Y	HA1
216	199	205 (HA1)	V	HA1
217	200	206 (HA1)	F	HA1
218	201	207 (HA1)	V	HA1
219	202	208 (HA1)	G	HA1
220	203	209 (HA1)	S	HA1
221	204	210 (HA1)	S	HA1
222	205	211 (HA1)	R	HA1
223	206	212 (HA1)	Y	HA1
224	207	213 (HA1)	S	HA1
225	208	214 (HA1)	K	HA1
226	209	215 (HA1)	K	HA1
227	210	216 (HA1)	F	HA1
228	211	217 (HA1)	K	HA1
229	212	218 (HA1)	P	HA1
230	213	219 (HA1)	E	HA1
231	214	220 (HA1)	I	HA1
232	215	221 (HA1)	A	HA1
233	216	222 (HA1)	I	HA1
234	217	223 (HA1)	R	HA1
235	218	224 (HA1)	P	HA1
236	219	225 (HA1)	K	HA1
237	220	226 (HA1)	V	HA1
238	221	227 (HA1)	R	HA1
239	222	228 (HA1)	D	HA1
240	223	229 (HA1)	Q	HA1
241	224	230 (HA1)	E	HA1
242	225	231 (HA1)	G	HA1
243	226	232 (HA1)	R	HA1
244	227	233 (HA1)	M	HA1
245	228	234 (HA1)	N	HA1

246	229	235 (HA1)	Y	HA1
247	230	236 (HA1)	Y	HA1
248	231	237 (HA1)	W	HA1
249	232	238 (HA1)	T	HA1
250	233	239 (HA1)	L	HA1
251	234	240 (HA1)	V	HA1
252	235	241 (HA1)	E	HA1
253	236	242 (HA1)	P	HA1
254	237	243 (HA1)	G	HA1
255	238	244 (HA1)	D	HA1
256	239	245 (HA1)	K	HA1
257	240	246 (HA1)	I	HA1
258	241	247 (HA1)	T	HA1
259	242	248 (HA1)	F	HA1
260	243	249 (HA1)	E	HA1
261	244	250 (HA1)	A	HA1
262	245	251 (HA1)	T	HA1
263	246	252 (HA1)	G	HA1
264	247	253 (HA1)	N	HA1
265	248	254 (HA1)	L	HA1
266	249	255 (HA1)	V	HA1
267	250	256 (HA1)	V	HA1
268	251	257 (HA1)	P	HA1
269	252	258 (HA1)	R	HA1
270	253	259 (HA1)	Y	HA1
271	254	260 (HA1)	A	HA1
272	255	261 (HA1)	F	HA1
273	256	262 (HA1)	A	HA1
274	257	263 (HA1)	M	HA1
275	258	264 (HA1)	E	HA1
276	259	265 (HA1)	R	HA1
277	260	266 (HA1)	N	HA1
278	261	267 (HA1)	A	HA1
279	262	268 (HA1)	G	HA1
280	263	269 (HA1)	S	HA1
281	264	270 (HA1)	G	HA1
282	265	271 (HA1)	I	HA1
283	266	272 (HA1)	I	HA1
284	267	273 (HA1)	I	HA1
285	268	274 (HA1)	S	HA1

286	269	275 (HA1)	D	HA1
287	270	276 (HA1)	T	HA1
288	271	277 (HA1)	P	HA1
289	272	278 (HA1)	V	HA1
290	273	279 (HA1)	H	HA1
291	274	280 (HA1)	D	HA1
292	275	281 (HA1)	C	HA1
293	276	282 (HA1)	N	HA1
294	277	283 (HA1)	T	HA1
295	278	284 (HA1)	T	HA1
296	279	285 (HA1)	C	HA1
297	280	286 (HA1)	Q	HA1
298	281	287 (HA1)	T	HA1
299	282	288 (HA1)	P	HA1
300	283	289 (HA1)	K	HA1
301	284	290 (HA1)	G	HA1
302	285	291 (HA1)	A	HA1
303	286	292 (HA1)	I	HA1
304	287	293 (HA1)	N	HA1
305	288	294 (HA1)	T	HA1
306	289	295 (HA1)	S	HA1
307	290	296 (HA1)	L	HA1
308	291	297 (HA1)	P	HA1
309	292	298 (HA1)	F	HA1
310	293	299 (HA1)	Q	HA1
311	294	300 (HA1)	N	HA1
312	295	301 (HA1)	I	HA1
313	296	302 (HA1)	H	HA1
314	297	303 (HA1)	P	HA1
315	298	304 (HA1)	I	HA1
316	299	305 (HA1)	T	HA1
317	300	306 (HA1)	I	HA1
318	301	307 (HA1)	G	HA1
319	302	308 (HA1)	K	HA1
320	303	309 (HA1)	C	HA1
321	304	310 (HA1)	P	HA1
322	305	311 (HA1)	K	HA1
323	306	312 (HA1)	Y	HA1
324	307	313 (HA1)	V	HA1
325	308	314 (HA1)	K	HA1

326	309	315 (HA1)	S	HA1
327	310	316 (HA1)	T	HA1
328	311	317 (HA1)	K	HA1
329	312	318 (HA1)	L	HA1
330	313	319 (HA1)	R	HA1
331	314	320 (HA1)	L	HA1
332	315	321 (HA1)	A	HA1
333	316	322 (HA1)	T	HA1
334	317	323 (HA1)	G	HA1
335	318	324 (HA1)	L	HA1
336	319	325 (HA1)	R	HA1
337	320	326 (HA1)	N	HA1
338	321	327 (HA1)	I	HA1
339	322	-	P	
340	323	-	S	
341	324	-	I	
342	325	-	Q	
343	326	-	S	
344	327	-	R	
345	328	1 (HA2)	G	HA2
346	329	2 (HA2)	L	HA2
347	330	3 (HA2)	F	HA2
348	331	4 (HA2)	G	HA2
349	332	5 (HA2)	A	HA2
350	333	6 (HA2)	I	HA2
351	334	7 (HA2)	A	HA2
352	335	8 (HA2)	G	HA2
353	336	9 (HA2)	F	HA2
354	337	10 (HA2)	I	HA2
355	338	11 (HA2)	E	HA2
356	339	12 (HA2)	G	HA2
357	340	13 (HA2)	G	HA2
358	341	14 (HA2)	W	HA2
359	342	15 (HA2)	T	HA2
360	343	16 (HA2)	G	HA2
361	344	17 (HA2)	M	HA2
362	345	18 (HA2)	V	HA2
363	346	19 (HA2)	D	HA2
364	347	20 (HA2)	G	HA2
365	348	21 (HA2)	W	HA2

366	349	22 (HA2)	Y	HA2
367	350	23 (HA2)	G	HA2
368	351	24 (HA2)	Y	HA2
369	352	25 (HA2)	H	HA2
370	353	26 (HA2)	H	HA2
371	354	27 (HA2)	Q	HA2
372	355	28 (HA2)	N	HA2
373	356	29 (HA2)	E	HA2
374	357	30 (HA2)	Q	HA2
375	358	31 (HA2)	G	HA2
376	359	32 (HA2)	S	HA2
377	360	33 (HA2)	G	HA2
378	361	34 (HA2)	Y	HA2
379	362	35 (HA2)	A	HA2
380	363	36 (HA2)	A	HA2
381	364	37 (HA2)	D	HA2
382	365	38 (HA2)	L	HA2
383	366	39 (HA2)	K	HA2
384	367	40 (HA2)	S	HA2
385	368	41 (HA2)	T	HA2
386	369	42 (HA2)	Q	HA2
387	370	43 (HA2)	N	HA2
388	371	44 (HA2)	A	HA2
389	372	45 (HA2)	I	HA2
390	373	46 (HA2)	D	HA2
391	374	47 (HA2)	E	HA2
392	375	48 (HA2)	I	HA2
393	376	49 (HA2)	T	HA2
394	377	50 (HA2)	N	HA2
395	378	51 (HA2)	K	HA2
396	379	52 (HA2)	V	HA2
397	380	53 (HA2)	N	HA2
398	381	54 (HA2)	S	HA2
399	382	55 (HA2)	V	HA2
400	383	56 (HA2)	I	HA2
401	384	57 (HA2)	E	HA2
402	385	58 (HA2)	K	HA2
403	386	59 (HA2)	M	HA2
404	387	60 (HA2)	N	HA2
405	388	61 (HA2)	T	HA2

406	389	62 (HA2)	Q	HA2
407	390	63 (HA2)	F	HA2
408	391	64 (HA2)	T	HA2
409	392	65 (HA2)	A	HA2
410	393	66 (HA2)	V	HA2
411	394	67 (HA2)	G	HA2
412	395	68 (HA2)	K	HA2
413	396	69 (HA2)	E	HA2
414	397	70 (HA2)	F	HA2
415	398	71 (HA2)	N	HA2
416	399	72 (HA2)	H	HA2
417	400	73 (HA2)	L	HA2
418	401	74 (HA2)	E	HA2
419	402	75 (HA2)	K	HA2
420	403	76 (HA2)	R	HA2
421	404	77 (HA2)	I	HA2
422	405	78 (HA2)	E	HA2
423	406	79 (HA2)	N	HA2
424	407	80 (HA2)	L	HA2
425	408	81 (HA2)	N	HA2
426	409	82 (HA2)	K	HA2
427	410	83 (HA2)	K	HA2
428	411	84 (HA2)	V	HA2
429	412	85 (HA2)	D	HA2
430	413	86 (HA2)	D	HA2
431	414	87 (HA2)	G	HA2
432	415	88 (HA2)	F	HA2
433	416	89 (HA2)	L	HA2
434	417	90 (HA2)	D	HA2
435	418	91 (HA2)	I	HA2
436	419	92 (HA2)	W	HA2
437	420	93 (HA2)	T	HA2
438	421	94 (HA2)	Y	HA2
439	422	95 (HA2)	N	HA2
440	423	96 (HA2)	A	HA2
441	424	97 (HA2)	E	HA2
442	425	98 (HA2)	L	HA2
443	426	99 (HA2)	L	HA2
444	427	100 (HA2)	V	HA2
445	428	101 (HA2)	L	HA2

446	429	102 (HA2)	L	HA2
447	430	103 (HA2)	E	HA2
448	431	104 (HA2)	N	HA2
449	432	105 (HA2)	E	HA2
450	433	106 (HA2)	R	HA2
451	434	107 (HA2)	T	HA2
452	435	108 (HA2)	L	HA2
453	436	109 (HA2)	D	HA2
454	437	110 (HA2)	Y	HA2
455	438	111 (HA2)	H	HA2
456	439	112 (HA2)	D	HA2
457	440	113 (HA2)	S	HA2
458	441	114 (HA2)	N	HA2
459	442	115 (HA2)	V	HA2
460	443	116 (HA2)	K	HA2
461	444	117 (HA2)	N	HA2
462	445	118 (HA2)	L	HA2
463	446	119 (HA2)	Y	HA2
464	447	120 (HA2)	E	HA2
465	448	121 (HA2)	K	HA2
466	449	122 (HA2)	V	HA2
467	450	123 (HA2)	R	HA2
468	451	124 (HA2)	S	HA2
469	452	125 (HA2)	Q	HA2
470	453	126 (HA2)	L	HA2
471	454	127 (HA2)	K	HA2
472	455	128 (HA2)	N	HA2
473	456	129 (HA2)	N	HA2
474	457	130 (HA2)	A	HA2
475	458	131 (HA2)	K	HA2
476	459	132 (HA2)	E	HA2
477	460	133 (HA2)	I	HA2
478	461	134 (HA2)	G	HA2
479	462	135 (HA2)	N	HA2
480	463	136 (HA2)	G	HA2
481	464	137 (HA2)	C	HA2
482	465	138 (HA2)	F	HA2
483	466	139 (HA2)	E	HA2
484	467	140 (HA2)	F	HA2
485	468	141 (HA2)	Y	HA2

486	469	142 (HA2)	H	HA2
487	470	143 (HA2)	K	HA2
488	471	144 (HA2)	C	HA2
489	472	145 (HA2)	D	HA2
490	473	146 (HA2)	N	HA2
491	474	147 (HA2)	T	HA2
492	475	148 (HA2)	C	HA2
493	476	149 (HA2)	M	HA2
494	477	150 (HA2)	E	HA2
495	478	151 (HA2)	S	HA2
496	479	152 (HA2)	V	HA2
497	480	153 (HA2)	K	HA2
498	481	154 (HA2)	N	HA2
499	482	155 (HA2)	G	HA2
500	483	156 (HA2)	T	HA2
501	484	157 (HA2)	Y	HA2
502	485	158 (HA2)	D	HA2
503	486	159 (HA2)	Y	HA2
504	487	160 (HA2)	P	HA2
505	488	161 (HA2)	K	HA2
506	489	162 (HA2)	Y	HA2
507	490	-	S	
508	491	-	E	
509	492	-	E	
510	493	-	A	
511	494	-	K	
512	495	-	L	
513	496	-	N	
514	497	-	R	
515	498	-	E	
516	499	-	E	
517	500	-	I	
518	501	-	D	
519	502	-	G	
520	503	-	V	Flexible Linker
521	504	-	K	Flexible Linker
522	505	-	L	Flexible Linker
523	506	-	E	Flexible Linker
524	507	-	S	Flexible Linker
525	508	-	T	Flexible Linker

526	509	-	R	Flexible Linker
527	510	-	I	Flexible Linker
528	511	-	Y	
529	512	-	Q	Transmembrane
530	513	-	I	Transmembrane
531	514	-	L	Transmembrane
532	515	-	A	Transmembrane
533	516	-	I	Transmembrane
534	517	-	Y	Transmembrane
535	518	-	S	Transmembrane
536	519	-	T	Transmembrane
537	520	-	V	Transmembrane
538	521	-	A	Transmembrane
539	522	-	S	Transmembrane
540	523	-	S	Transmembrane
541	524	-	L	Transmembrane
542	525	-	V	Transmembrane
543	526	-	L	Transmembrane
544	527	-	V	Transmembrane
545	528	-	V	Transmembrane
546	529	-	S	Transmembrane
547	530	-	L	Transmembrane
548	531	-	G	Transmembrane
549	532	-	A	Transmembrane
550	533	-	I	Transmembrane
551	534	-	S	Transmembrane
552	535	-	F	Transmembrane
553	536	-	W	Transmembrane
554	537	-	M	Transmembrane
555	538	-	C	
556	539	-	S	
557	540	-	N	
558	541	-	G	
559	542	-	S	
560	543	-	L	
561	544	-	Q	
562	545	-	C	
563	546	-	R	
564	547	-	I	
565	548	-	C	

566	549	-		I	

Table S4. Comparison of multiple HA numbering schemes using a H3 (A/Aichi/2/1968(H3N2)) template. Template for FLU DB numbering is H3. Flexible linker and transmembrane domain were located by aligning to A/duck/Alberta/35/76(H1N1), PDB ID: 6HJR.

position on CDS	FLU DB H3 numbering	structure-based H3 numbering (4HMG)	Residue	Annotation
1	-	-	M	Signal peptide
2	-	-	K	Signal peptide
3	-	-	T	Signal peptide
4	-	-	I	Signal peptide
5	-	-	I	Signal peptide
6	-	-	A	Signal peptide
7	-	-	L	Signal peptide
8	-	-	S	Signal peptide
9	-	-	Y	Signal peptide
10	-	-	I	Signal peptide
11	-	-	L	Signal peptide
12	-	-	C	Signal peptide
13	-	-	L	Signal peptide
14	-	-	V	Signal peptide
15	-	-	F	Signal peptide
16	-	-	A	Signal peptide
17	1	1 (HA1)	Q	HA1
18	2	2 (HA1)	D	HA1
19	3	3 (HA1)	L	HA1
20	4	4 (HA1)	P	HA1
21	5	5 (HA1)	G	HA1
22	6	6 (HA1)	N	HA1
23	7	7 (HA1)	D	HA1
24	8	8 (HA1)	N	HA1
25	9	9 (HA1)	S	HA1
26	10	10 (HA1)	T	HA1
27	11	11 (HA1)	A	HA1
28	12	12 (HA1)	T	HA1
29	13	13 (HA1)	L	HA1
30	14	14 (HA1)	C	HA1
31	15	15 (HA1)	L	HA1
32	16	16 (HA1)	G	HA1
33	17	17 (HA1)	H	HA1
34	18	18 (HA1)	H	HA1
35	19	19 (HA1)	A	HA1
36	20	20 (HA1)	V	HA1
37	21	21 (HA1)	P	HA1
38	22	22 (HA1)	N	HA1
39	23	23 (HA1)	G	HA1
40	24	24 (HA1)	T	HA1
41	25	25 (HA1)	L	HA1
42	26	26 (HA1)	V	HA1
43	27	27 (HA1)	K	HA1
44	28	28 (HA1)	T	HA1
45	29	29 (HA1)	I	HA1
46	30	30 (HA1)	T	HA1

47	31	31 (HA1)	D	HA1
48	32	32 (HA1)	D	HA1
49	33	33 (HA1)	Q	HA1
50	34	34 (HA1)	I	HA1
51	35	35 (HA1)	E	HA1
52	36	36 (HA1)	V	HA1
53	37	37 (HA1)	T	HA1
54	38	38 (HA1)	N	HA1
55	39	39 (HA1)	A	HA1
56	40	40 (HA1)	T	HA1
57	41	41 (HA1)	E	HA1
58	42	42 (HA1)	L	HA1
59	43	43 (HA1)	V	HA1
60	44	44 (HA1)	Q	HA1
61	45	45 (HA1)	S	HA1
62	46	46 (HA1)	S	HA1
63	47	47 (HA1)	S	HA1
64	48	48 (HA1)	T	HA1
65	49	49 (HA1)	G	HA1
66	50	50 (HA1)	K	HA1
67	51	51 (HA1)	I	HA1
68	52	52 (HA1)	C	HA1
69	53	53 (HA1)	N	HA1
70	54	54 (HA1)	N	HA1
71	55	55 (HA1)	P	HA1
72	56	56 (HA1)	H	HA1
73	57	57 (HA1)	R	HA1
74	58	58 (HA1)	I	HA1
75	59	59 (HA1)	L	HA1
76	60	60 (HA1)	D	HA1
77	61	61 (HA1)	G	HA1
78	62	62 (HA1)	I	HA1
79	63	63 (HA1)	D	HA1
80	64	64 (HA1)	C	HA1
81	65	65 (HA1)	T	HA1
82	66	66 (HA1)	L	HA1
83	67	67 (HA1)	I	HA1
84	68	68 (HA1)	D	HA1
85	69	69 (HA1)	A	HA1
86	70	70 (HA1)	L	HA1
87	71	71 (HA1)	L	HA1
88	72	72 (HA1)	G	HA1
89	73	73 (HA1)	D	HA1
90	74	74 (HA1)	P	HA1
91	75	75 (HA1)	H	HA1
92	76	76 (HA1)	C	HA1
93	77	77 (HA1)	D	HA1
94	78	78 (HA1)	V	HA1
95	79	79 (HA1)	F	HA1
96	80	80 (HA1)	Q	HA1
97	81	81 (HA1)	N	HA1
98	82	82 (HA1)	E	HA1
99	83	83 (HA1)	T	HA1
100	84	84 (HA1)	W	HA1

101	85	85 (HA1)	D	HA1
102	86	86 (HA1)	L	HA1
103	87	87 (HA1)	F	HA1
104	88	88 (HA1)	V	HA1
105	89	89 (HA1)	E	HA1
106	90	90 (HA1)	R	HA1
107	91	91 (HA1)	S	HA1
108	92	92 (HA1)	K	HA1
109	93	93 (HA1)	A	HA1
110	94	94 (HA1)	F	HA1
111	95	95 (HA1)	S	HA1
112	96	96 (HA1)	N	HA1
113	97	97 (HA1)	C	HA1
114	98	98 (HA1)	Y	HA1
115	99	99 (HA1)	P	HA1
116	100	100 (HA1)	Y	HA1
117	101	101 (HA1)	D	HA1
118	102	102 (HA1)	V	HA1
119	103	103 (HA1)	P	HA1
120	104	104 (HA1)	D	HA1
121	105	105 (HA1)	Y	HA1
122	106	106 (HA1)	A	HA1
123	107	107 (HA1)	S	HA1
124	108	108 (HA1)	L	HA1
125	109	109 (HA1)	R	HA1
126	110	110 (HA1)	S	HA1
127	111	111 (HA1)	L	HA1
128	112	112 (HA1)	V	HA1
129	113	113 (HA1)	A	HA1
130	114	114 (HA1)	S	HA1
131	115	115 (HA1)	S	HA1
132	116	116 (HA1)	G	HA1
133	117	117 (HA1)	T	HA1
134	118	118 (HA1)	L	HA1
135	119	119 (HA1)	E	HA1
136	120	120 (HA1)	F	HA1
137	121	121 (HA1)	I	HA1
138	122	122 (HA1)	T	HA1
139	123	123 (HA1)	E	HA1
140	124	124 (HA1)	G	HA1
141	125	125 (HA1)	F	HA1
142	126	126 (HA1)	T	HA1
143	127	127 (HA1)	W	HA1
144	128	128 (HA1)	T	HA1
145	129	129 (HA1)	G	HA1
146	130	130 (HA1)	V	HA1
147	131	131 (HA1)	T	HA1
148	132	132 (HA1)	Q	HA1
149	133	133 (HA1)	N	HA1
150	134	134 (HA1)	G	HA1
151	135	135 (HA1)	G	HA1
152	136	136 (HA1)	S	HA1
153	137	137 (HA1)	N	HA1
154	138	138 (HA1)	A	HA1

155	139	139 (HA1)	C	HA1
156	140	140 (HA1)	K	HA1
157	141	141 (HA1)	R	HA1
158	142	142 (HA1)	G	HA1
159	143	143 (HA1)	P	HA1
160	144	144 (HA1)	G	HA1
161	145	145 (HA1)	S	HA1
162	146	146 (HA1)	G	HA1
163	147	147 (HA1)	F	HA1
164	148	148 (HA1)	F	HA1
165	149	149 (HA1)	S	HA1
166	150	150 (HA1)	R	HA1
167	151	151 (HA1)	L	HA1
168	152	152 (HA1)	N	HA1
169	153	153 (HA1)	W	HA1
170	154	154 (HA1)	L	HA1
171	155	155 (HA1)	T	HA1
172	156	156 (HA1)	K	HA1
173	157	157 (HA1)	S	HA1
174	158	158 (HA1)	G	HA1
175	159	159 (HA1)	S	HA1
176	160	160 (HA1)	T	HA1
177	161	161 (HA1)	Y	HA1
178	162	162 (HA1)	P	HA1
179	163	163 (HA1)	V	HA1
180	164	164 (HA1)	L	HA1
181	165	165 (HA1)	N	HA1
182	166	166 (HA1)	V	HA1
183	167	167 (HA1)	T	HA1
184	168	168 (HA1)	M	HA1
185	169	169 (HA1)	P	HA1
186	170	170 (HA1)	N	HA1
187	171	171 (HA1)	N	HA1
188	172	172 (HA1)	D	HA1
189	173	173 (HA1)	N	HA1
190	174	174 (HA1)	F	HA1
191	175	175 (HA1)	D	HA1
192	176	176 (HA1)	K	HA1
193	177	177 (HA1)	L	HA1
194	178	178 (HA1)	Y	HA1
195	179	179 (HA1)	I	HA1
196	180	180 (HA1)	W	HA1
197	181	181 (HA1)	G	HA1
198	182	182 (HA1)	I	HA1
199	183	183 (HA1)	H	HA1
200	184	184 (HA1)	H	HA1
201	185	185 (HA1)	P	HA1
202	186	186 (HA1)	S	HA1
203	187	187 (HA1)	T	HA1
204	188	188 (HA1)	N	HA1
205	189	189 (HA1)	Q	HA1
206	190	190 (HA1)	E	HA1
207	191	191 (HA1)	Q	HA1
208	192	192 (HA1)	T	HA1

209	193	193 (HA1)	S	HA1
210	194	194 (HA1)	L	HA1
211	195	195 (HA1)	Y	HA1
212	196	196 (HA1)	V	HA1
213	197	197 (HA1)	Q	HA1
214	198	198 (HA1)	A	HA1
215	199	199 (HA1)	S	HA1
216	200	200 (HA1)	G	HA1
217	201	201 (HA1)	R	HA1
218	202	202 (HA1)	V	HA1
219	203	203 (HA1)	T	HA1
220	204	204 (HA1)	V	HA1
221	205	205 (HA1)	S	HA1
222	206	206 (HA1)	T	HA1
223	207	207 (HA1)	R	HA1
224	208	208 (HA1)	R	HA1
225	209	209 (HA1)	S	HA1
226	210	210 (HA1)	Q	HA1
227	211	211 (HA1)	Q	HA1
228	212	212 (HA1)	T	HA1
229	213	213 (HA1)	I	HA1
230	214	214 (HA1)	I	HA1
231	215	215 (HA1)	P	HA1
232	216	216 (HA1)	N	HA1
233	217	217 (HA1)	I	HA1
234	218	218 (HA1)	G	HA1
235	219	219 (HA1)	S	HA1
236	220	220 (HA1)	R	HA1
237	221	221 (HA1)	P	HA1
238	222	222 (HA1)	W	HA1
239	223	223 (HA1)	V	HA1
240	224	224 (HA1)	R	HA1
241	225	225 (HA1)	G	HA1
242	226	226 (HA1)	L	HA1
243	227	227 (HA1)	S	HA1
244	228	228 (HA1)	S	HA1
245	229	229 (HA1)	R	HA1
246	230	230 (HA1)	I	HA1
247	231	231 (HA1)	S	HA1
248	232	232 (HA1)	I	HA1
249	233	233 (HA1)	Y	HA1
250	234	234 (HA1)	W	HA1
251	235	235 (HA1)	T	HA1
252	236	236 (HA1)	I	HA1
253	237	237 (HA1)	V	HA1
254	238	238 (HA1)	K	HA1
255	239	239 (HA1)	P	HA1
256	240	240 (HA1)	G	HA1
257	241	241 (HA1)	D	HA1
258	242	242 (HA1)	V	HA1
259	243	243 (HA1)	L	HA1
260	244	244 (HA1)	V	HA1
261	245	245 (HA1)	I	HA1
262	246	246 (HA1)	N	HA1

263	247	247 (HA1)	S	HA1
264	248	248 (HA1)	N	HA1
265	249	249 (HA1)	G	HA1
266	250	250 (HA1)	N	HA1
267	251	251 (HA1)	L	HA1
268	252	252 (HA1)	I	HA1
269	253	253 (HA1)	A	HA1
270	254	254 (HA1)	P	HA1
271	255	255 (HA1)	R	HA1
272	256	256 (HA1)	G	HA1
273	257	257 (HA1)	Y	HA1
274	258	258 (HA1)	F	HA1
275	259	259 (HA1)	K	HA1
276	260	260 (HA1)	M	HA1
277	261	261 (HA1)	R	HA1
278	262	262 (HA1)	T	HA1
279	263	263 (HA1)	G	HA1
280	264	264 (HA1)	K	HA1
281	265	265 (HA1)	S	HA1
282	266	266 (HA1)	S	HA1
283	267	267 (HA1)	I	HA1
284	268	268 (HA1)	M	HA1
285	269	269 (HA1)	R	HA1
286	270	270 (HA1)	S	HA1
287	271	271 (HA1)	D	HA1
288	272	272 (HA1)	A	HA1
289	273	273 (HA1)	P	HA1
290	274	274 (HA1)	I	HA1
291	275	275 (HA1)	D	HA1
292	276	276 (HA1)	T	HA1
293	277	277 (HA1)	C	HA1
294	278	278 (HA1)	I	HA1
295	279	279 (HA1)	S	HA1
296	280	280 (HA1)	E	HA1
297	281	281 (HA1)	C	HA1
298	282	282 (HA1)	I	HA1
299	283	283 (HA1)	T	HA1
300	284	284 (HA1)	P	HA1
301	285	285 (HA1)	N	HA1
302	286	286 (HA1)	G	HA1
303	287	287 (HA1)	S	HA1
304	288	288 (HA1)	I	HA1
305	289	289 (HA1)	P	HA1
306	290	290 (HA1)	N	HA1
307	291	291 (HA1)	D	HA1
308	292	292 (HA1)	K	HA1
309	293	293 (HA1)	P	HA1
310	294	294 (HA1)	F	HA1
311	295	295 (HA1)	Q	HA1
312	296	296 (HA1)	N	HA1
313	297	297 (HA1)	V	HA1
314	298	298 (HA1)	N	HA1
315	299	299 (HA1)	K	HA1
316	300	300 (HA1)	I	HA1

317	301	301 (HA1)	T	HA1
318	302	302 (HA1)	Y	HA1
319	303	303 (HA1)	G	HA1
320	304	304 (HA1)	A	HA1
321	305	305 (HA1)	C	HA1
322	306	306 (HA1)	P	HA1
323	307	307 (HA1)	K	HA1
324	308	308 (HA1)	Y	HA1
325	309	309 (HA1)	V	HA1
326	310	310 (HA1)	K	HA1
327	311	311 (HA1)	Q	HA1
328	312	312 (HA1)	N	HA1
329	313	313 (HA1)	T	HA1
330	314	314 (HA1)	L	HA1
331	315	315 (HA1)	K	HA1
332	316	316 (HA1)	L	HA1
333	317	317 (HA1)	A	HA1
334	318	318 (HA1)	T	HA1
335	319	319 (HA1)	G	HA1
336	320	320 (HA1)	M	HA1
337	321	321 (HA1)	R	HA1
338	322	322 (HA1)	N	HA1
339	323	323 (HA1)	V	HA1
340	324	324 (HA1)	P	HA1
341	325	325 (HA1)	E	HA1
342	326	326 (HA1)	K	HA1
343	327	327 (HA1)	Q	HA1
344	328	328 (HA1)	T	HA1
345	329	-	R	
346	330	1 (HA2)	G	HA2
347	331	2 (HA2)	L	HA2
348	332	3 (HA2)	F	HA2
349	333	4 (HA2)	G	HA2
350	334	5 (HA2)	A	HA2
351	335	6 (HA2)	I	HA2
352	336	7 (HA2)	A	HA2
353	337	8 (HA2)	G	HA2
354	338	9 (HA2)	F	HA2
355	339	1 (HA2)0	I	HA2
356	340	11 (HA2)	E	HA2
357	341	12 (HA2)	N	HA2
358	342	13 (HA2)	G	HA2
359	343	14 (HA2)	W	HA2
360	344	15 (HA2)	E	HA2
361	345	16 (HA2)	G	HA2
362	346	17 (HA2)	M	HA2
363	347	18 (HA2)	I	HA2
364	348	19 (HA2)	D	HA2
365	349	20 (HA2)	G	HA2
366	350	21 (HA2)	W	HA2
367	351	22 (HA2)	Y	HA2
368	352	23 (HA2)	G	HA2
369	353	24 (HA2)	F	HA2
370	354	25 (HA2)	R	HA2

371	355	26 (HA2)	H	HA2
372	356	27 (HA2)	Q	HA2
373	357	28 (HA2)	N	HA2
374	358	29 (HA2)	S	HA2
375	359	30 (HA2)	E	HA2
376	360	31 (HA2)	G	HA2
377	361	32 (HA2)	T	HA2
378	362	33 (HA2)	G	HA2
379	363	34 (HA2)	Q	HA2
380	364	35 (HA2)	A	HA2
381	365	36 (HA2)	A	HA2
382	366	37 (HA2)	D	HA2
383	367	38 (HA2)	L	HA2
384	368	39 (HA2)	K	HA2
385	369	40 (HA2)	S	HA2
386	370	41 (HA2)	T	HA2
387	371	42 (HA2)	Q	HA2
388	372	43 (HA2)	A	HA2
389	373	44 (HA2)	A	HA2
390	374	45 (HA2)	I	HA2
391	375	46 (HA2)	D	HA2
392	376	47 (HA2)	Q	HA2
393	377	48 (HA2)	I	HA2
394	378	49 (HA2)	N	HA2
395	379	50 (HA2)	G	HA2
396	380	51 (HA2)	K	HA2
397	381	52 (HA2)	L	HA2
398	382	53 (HA2)	N	HA2
399	383	54 (HA2)	R	HA2
400	384	55 (HA2)	V	HA2
401	385	56 (HA2)	I	HA2
402	386	57 (HA2)	E	HA2
403	387	58 (HA2)	K	HA2
404	388	59 (HA2)	T	HA2
405	389	60 (HA2)	N	HA2
406	390	61 (HA2)	E	HA2
407	391	62 (HA2)	K	HA2
408	392	63 (HA2)	F	HA2
409	393	64 (HA2)	H	HA2
410	394	65 (HA2)	Q	HA2
411	395	66 (HA2)	I	HA2
412	396	67 (HA2)	E	HA2
413	397	68 (HA2)	K	HA2
414	398	69 (HA2)	E	HA2
415	399	70 (HA2)	F	HA2
416	400	71 (HA2)	S	HA2
417	401	72 (HA2)	E	HA2
418	402	73 (HA2)	V	HA2
419	403	74 (HA2)	E	HA2
420	404	75 (HA2)	G	HA2
421	405	76 (HA2)	R	HA2
422	406	77 (HA2)	I	HA2
423	407	78 (HA2)	Q	HA2
424	408	79 (HA2)	D	HA2

425	409	80 (HA2)	L	HA2
426	410	81 (HA2)	E	HA2
427	411	82 (HA2)	K	HA2
428	412	83 (HA2)	Y	HA2
429	413	84 (HA2)	V	HA2
430	414	85 (HA2)	E	HA2
431	415	86 (HA2)	D	HA2
432	416	87 (HA2)	T	HA2
433	417	88 (HA2)	K	HA2
434	418	89 (HA2)	I	HA2
435	419	90 (HA2)	D	HA2
436	420	91 (HA2)	L	HA2
437	421	92 (HA2)	W	HA2
438	422	93 (HA2)	S	HA2
439	423	94 (HA2)	Y	HA2
440	424	95 (HA2)	N	HA2
441	425	96 (HA2)	A	HA2
442	426	97 (HA2)	E	HA2
443	427	98 (HA2)	L	HA2
444	428	99 (HA2)	L	HA2
445	429	10 (HA2)0	V	HA2
446	430	101 (HA2)	A	HA2
447	431	102 (HA2)	L	HA2
448	432	103 (HA2)	E	HA2
449	433	104 (HA2)	N	HA2
450	434	105 (HA2)	Q	HA2
451	435	106 (HA2)	H	HA2
452	436	107 (HA2)	T	HA2
453	437	108 (HA2)	I	HA2
454	438	109 (HA2)	D	HA2
455	439	110 (HA2)	L	HA2
456	440	111 (HA2)	T	HA2
457	441	112 (HA2)	D	HA2
458	442	113 (HA2)	S	HA2
459	443	114 (HA2)	E	HA2
460	444	115 (HA2)	M	HA2
461	445	116 (HA2)	N	HA2
462	446	117 (HA2)	K	HA2
463	447	118 (HA2)	L	HA2
464	448	119 (HA2)	F	HA2
465	449	120 (HA2)	E	HA2
466	450	121 (HA2)	K	HA2
467	451	122 (HA2)	T	HA2
468	452	123 (HA2)	R	HA2
469	453	124 (HA2)	R	HA2
470	454	125 (HA2)	Q	HA2
471	455	126 (HA2)	L	HA2
472	456	127 (HA2)	R	HA2
473	457	128 (HA2)	E	HA2
474	458	129 (HA2)	N	HA2
475	459	130 (HA2)	A	HA2
476	460	131 (HA2)	E	HA2
477	461	132 (HA2)	E	HA2
478	462	133 (HA2)	M	HA2

479	463	134 (HA2)	G	HA2
480	464	135 (HA2)	N	HA2
481	465	136 (HA2)	G	HA2
482	466	137 (HA2)	C	HA2
483	467	138 (HA2)	F	HA2
484	468	139 (HA2)	K	HA2
485	469	140 (HA2)	I	HA2
486	470	141 (HA2)	Y	HA2
487	471	142 (HA2)	H	HA2
488	472	143 (HA2)	K	HA2
489	473	144 (HA2)	C	HA2
490	474	145 (HA2)	D	HA2
491	475	146 (HA2)	N	HA2
492	476	147 (HA2)	A	HA2
493	477	148 (HA2)	C	HA2
494	478	149 (HA2)	I	HA2
495	479	150 (HA2)	E	HA2
496	480	151 (HA2)	S	HA2
497	481	152 (HA2)	I	HA2
498	482	153 (HA2)	R	HA2
499	483	154 (HA2)	N	HA2
500	484	155 (HA2)	G	HA2
501	485	156 (HA2)	T	HA2
502	486	157 (HA2)	Y	HA2
503	487	158 (HA2)	D	HA2
504	488	159 (HA2)	H	HA2
505	489	160 (HA2)	D	HA2
506	490	161 (HA2)	V	HA2
507	491	162 (HA2)	Y	HA2
508	492	163 (HA2)	R	HA2
509	493	164 (HA2)	D	HA2
510	494	165 (HA2)	E	HA2
511	495	166 (HA2)	A	HA2
512	496	167 (HA2)	L	HA2
513	497	168 (HA2)	N	HA2
514	498	169 (HA2)	N	HA2
515	499	170 (HA2)	R	HA2
516	500	171 (HA2)	F	HA2
517	501	172 (HA2)	Q	HA2
518	502	173 (HA2)	I	HA2
519	503	174 (HA2)	K	HA2
520	504	175 (HA2)	G	HA2
521	505	-	V	Flexible Linker
522	506	-	E	Flexible Linker
523	507	-	L	Flexible Linker
524	508	-	K	Flexible Linker
525	509	-	S	Flexible Linker
526	510	-	G	Flexible Linker
527	511	-	Y	Flexible Linker
528	512	-	K	Flexible Linker
529	513	-	D	
530	514	-	W	Transmembrane
531	515	-	I	Transmembrane
532	516	-	L	Transmembrane

533	517	-	W	Transmembrane
534	518	-	I	Transmembrane
535	519	-	S	Transmembrane
536	520	-	F	Transmembrane
537	521	-	A	Transmembrane
538	522	-	I	Transmembrane
539	523	-	S	Transmembrane
540	524	-	C	Transmembrane
541	525	-	F	Transmembrane
542	526	-	L	Transmembrane
543	527	-	L	Transmembrane
544	528	-	C	Transmembrane
545	529	-	V	Transmembrane
546	530	-	V	Transmembrane
547	531	-	L	Transmembrane
548	532	-	L	Transmembrane
549	533	-	G	Transmembrane
550	534	-	F	Transmembrane
551	535	-	I	Transmembrane
552	536	-	M	Transmembrane
553	537	-	W	Transmembrane
554	538	-	A	Transmembrane
555	539	-	C	
556	540	-	Q	
557	541	-	R	
558	542	-	G	
559	543	-	N	
560	544	-	I	
561	545	-	R	
562	546	-	C	
563	547	-	N	
564	548	-	I	
565	549	-	C	
566	550	-	I	