Recognition of Dual Targets by Molecular Beacon-Based Sensor: Subtyping of Influenza A Virus

Chun-Ching Lee[1], Yu-Chieh Liao[2], Yu-Hsuan Lai[1],and Min-Chieh Chuang[1,*]
[1] Department of Chemistry, Tunghai University, Taichung, Taiwan
[2] Institute of Population Health Science, National Health Research Institutes, Zhunan, Miaoli County, Taiwan
* To whom correspondence should be addressed. Tel: +886-4-2359-0121 ext. 32218; Fax: +886-4-2359-0426; Email: mcchuang@thu.edu.tw

Table S-1. Oligonucleotides used in this study.

Designation	Sequence and modification (5'→3')	Notes
A2	CAT CAT CAC TAC AGA GGA GCT ATC ATG ATT	Assistant strand for (1,1)
A3	ATC ATC ACTACA GAT GAG CTA TCA TGA TTC	Assistant strand for (2,0)
A4	CAT CAT CAC TAC AGA GAG CTA TCA TGA TTC	Assistant strand for (2,1)
A7	CCA TCA GGC CAT GAC ATG ATT GCC AGT GC	Assistant strand for H5- and N2-specific sequences
A8	ATC ATC ACT ACA GAT TGG AGC TAT CAT GAT	Assistant strand for (0,0)
A9	CCA TCA TCA CTA CAG GGA GCT ATC ATG ATT	Assistant strand for (1,2)
H	GCA CTG GCA ATC ATG ATA GCT GGT CTA TCT TTT TGG TGA AGC	The target bearing H5-specific sequence
MB1	FAM-CCG GCA ATG ATT GAT CGT TAC CCA AAT AGT ATG CCG G-Dabcyl	Molecular beacon 1
MB2	FAM-CCA CAG GGT AAA GAT AGA CCA GCT GCC TGT TCC ATA CCC TGT GG-Dabcyl	Molecular beacon 2
N	TAT GGA ACA GGC TCA TGG CCT GAT GGG GCG A	The target bearing N2-specific sequence
TE	ACT CAA ACG ATC AAT CAT TGC	Target sequence E
TF	GCA TAC TAT TTG GGT CTA TC	Target sequence F
TG	AAT GAA TCA TGA TAG CTC CAA ACG ATC AAT CAT T	Target sequence G
TI	ATA CTA TTT GGG TAT CTG TAG TGA TGA TGG GGC	Target sequence I
SH5	GCA GCG AGT TCC CTA GTA CTG	Sense primer for H5
ASH5	TGA CCC ATT GGA ACA CAT CCA G	Antisense primer for H5
SN2	TTG TGG CAC TTC AGG TAC TTA TG	Sense primer for N2
ASN2	ATA GGC ATG AAG TTG ATA TTC GCC C	Antisense primer for N2

Asymmetric Polymerase Chain Reaction (aPCR)

The sense and antisense primers (Table S-1) for amplification of H5 and N2 genes from the cDNA of A/duck/Taiwan/DV30-2/2005 (H5N2) were designed by the Tm Calculator of NCBI primer-BLAST. The aPCR product of H5 segment was generated from a 25 μL PCR mixture containing 1X Phusion HF buffer, 200μM dNTP, 1 μM SH5, 0.1 μM ASH5, 1 μg viral genome (cDNA), and 0.5 units of Phusion High-Fidelity DNA polymerase. Reactions were performed with the MJ Mini Thermal Cycler (Bio-Rad, Hercules, CA). The amplification cycles include 1 cycle of 30 sec at 98 °C followed by 50 cycles of 98 °C for 10 sec, 62.2 °C for 20 sec, and 72 °C for 15 sec, and was finally cooled down at 4 °C for 30 sec. For the aPCR of N2 gene, the 25 μL PCR mixture composed of 1X Phusion HF buffer, 200 μM dNTP, 1 μM SN2, 0.2 μM ASN2, 1μg viral genome (cDNA), and 0.5 units of Phusion High-Fidelity DNA polymerase. The amplification cycles include 1 cycle of 30 sec at 98 °C followed by 50 cycles of 98 °C for 10 sec, 60.7 °C for 20 sec, and 72 °C for 15 sec, and was finally cooled down at 4 °C for 30 sec.

Viral samples and reverse transcription

Genuine sample derived from influenza virus particles, the strain A/duck/Taiwan/DV30-2/2005 (H5N2), was kindly supplied by the ID Public Health Lab (led by Professor Chwan-Chuen King) in College of Public Heath, National Taiwan University. The Genbank accession number is CY110933 for HA and CY110935 for NA gene. Viral RNA (vRNA) was extracted using QIAamp Viral RNA Mini Kit (Qiagen, Limburg, Netherlands) according to manufacturer's instructions. The extracted RNA were then reverse transcribed to cDNA using an universal primer Uni12, which is complementary to 12 conserved contiguous nucleotides at the 3'-end of influenza vRNA.[1] Prior to reverse transcription, mixture of vRNA, Uni12 primer, and dNTP was heated at 65 °C for 5 min, followed by an incubation on ice for 5 min. Subsequently, cDNA were generated in a solution (20 μL) composed of 5 μL vRNA, 500 μM dNTP, 500 nM Uni12 primer, 10 U Superscript™ III Reverse Transcriptase (Invitrogen, Carlsbad, CA), 10 mM DTT, 2U RNaseOUT™ Recombinant Ribonuclease Inhibitor (Invotrogen) in the first-strand buffer. The solution was subsequently incubated at 50 °C for 1 hr and the reaction was inactivated at 70 °C for 15-min. Eventually the cDNA products were purified with QIAquick PCR Purification Kit (Qiagen).

Equilibrium analysis

Equilibrium constants for the hybridization between the MB and its target hybrids were evaluated based on the method described below. In a solution containing MB and target DNA (T), there existed at least three distinct states: MB-T hybrid (phase 1, denoted as MB·T), stem-loop folded MB free from the target (phase 2, MB_{closed} + T), and random coiled MB free from the target (phase 3, MB_{open} + T). The reaction can be presented as:

$$\mathrm{MB{\cdot}T} \overset{K_{12}}{\rightleftharpoons} \mathrm{MB_{closed}} + \mathrm{T} \overset{K_{23}}{\rightleftharpoons} \mathrm{MB_{open}} + \mathrm{T} \qquad (1)$$

The fluorescence (F) obtained from the solution at a given temperature is the sum of the molecular beacons in each of the three states, and can be expressed as:

$$F = \alpha \frac{[\mathrm{MBT}]}{\mathrm{MB_0}} + \beta \frac{[\mathrm{MB_{closed}}]}{\mathrm{MB_0}} + \gamma \frac{[\mathrm{MB_{open}}]}{\mathrm{MB_0}} \qquad (2)$$

Where α, β, and γ are the characteristic fluorescence intensities of the molecular beacon in each state, and MB_0 is the total concentration of molecular beacon (= [MBT] + [MB_{closed}] + [MB_{open}]).Based on the basic definition of K_{12} and K_{23}:

$$K_{23} = \frac{[\mathrm{MB_{open}}]}{[\mathrm{MB_{closed}}]} \qquad (3)$$

$$K_{12} = \frac{[\mathrm{MB_{closed}}][\mathrm{T}]}{[\mathrm{MBT}]} \qquad (4)$$

When the total concentration of the target, T_0, is much greater than the total concentration of the molecular beacon, MB_0, T_0 can be substituted for [T]. The fraction of molecular beacons illustrated in eq. 2 can be expressed in terms of K_{12} and K_{23}; furthermore, eq. 1 can be presented as:

$$F = \frac{\alpha \mathrm{T_0} + \beta K_{12} + \gamma K_{12} K_{23}}{\mathrm{T_0} + K_{12} + K_{12} K_{23}} \qquad (5)$$

A rearrangement of eq. 5 results in:

$$K_{12} = \frac{(\alpha - F)\mathrm{T_0}}{(F-\beta) + (F-\gamma)K_{23}} \qquad (6)$$

Also by taking T_0 = 0 (as a circumstance in absence of targets), eq. 5 can be rearranged to

give:

$$K_{23} = \frac{\theta - \beta}{\gamma - \theta} \qquad (7)$$

In the evaluation of K_{12} and K_{23} based on eq. 6 and 7, F is the fluorescence intensity as a function of temperature obtained from the thermal denaturation profile. The value of β is the fluorescence of the molecular beacon solution (in absence of targets) acquired at 15 °C. The value of α and γ is the fluorescence measured at 15 and 85 °C, respectively, in the presence of targets.

Kinetic analysis

The unfolding kinetics of MB upon hybridization with target DNAs was analyzed:

$$\mathrm{MB} + \mathrm{T} \rightleftharpoons \mathrm{MB{\cdot}T}, \qquad (8)$$

$$\frac{\mathrm{d[MB{\cdot}T]}}{\mathrm{dt}} = k_1[\mathrm{MB}][\mathrm{T}] - k_2[\mathrm{MB{\cdot}T}] \qquad (9)$$

where k_1 is the opening rate constant, and k_2 is the closing rate constant of MB upon hybridization. Assuming that fluorescence could be normalized as following,

$$\frac{[\mathrm{MB{\cdot}T}]_T}{[\mathrm{MB{\cdot}T}]_{eq}} = \frac{F(T)-F_0}{F_{eq}-F_0} = F_n \qquad (10)$$

in which F_{eq} is the fluorescence of the system when $T \rightarrow \infty$ and F_0 is the initial fluorescence intensity. After solving the above equation and rearranging the formula with normalized fluorescence $\mathrm{F_n}$, an exponential function can be derived:

$$\frac{1-F_n}{1-\rho F_n} = e^{-\omega k_1 t} \qquad (11)$$

where $\rho = ([\mathrm{MB{\cdot}T}]_{eq})^2/\mathrm{MB_0T_0}$, $[\mathrm{MB{\cdot}T}]_{eq} = (\mathrm{MB_0} + \mathrm{T_0} + K_{12} - \omega)/2$,

$\omega = \sqrt{(\mathrm{MB_0} + \mathrm{T_0} + K_{12})^2 - 4\mathrm{MB_0T_0}}$, and $K_{12} = k_2/k_1$.

The unfolding rate constant k_1 could be derived by fitting the normalized fluorescence data to an alternative form of equation 11 (after taking natural log of it), resulting in a linear plot with a slope k_1,

$$\frac{1}{\omega} ln\left(\frac{1-F_n}{1-\rho F_n)}\right) = -k_1 t \qquad (12)$$

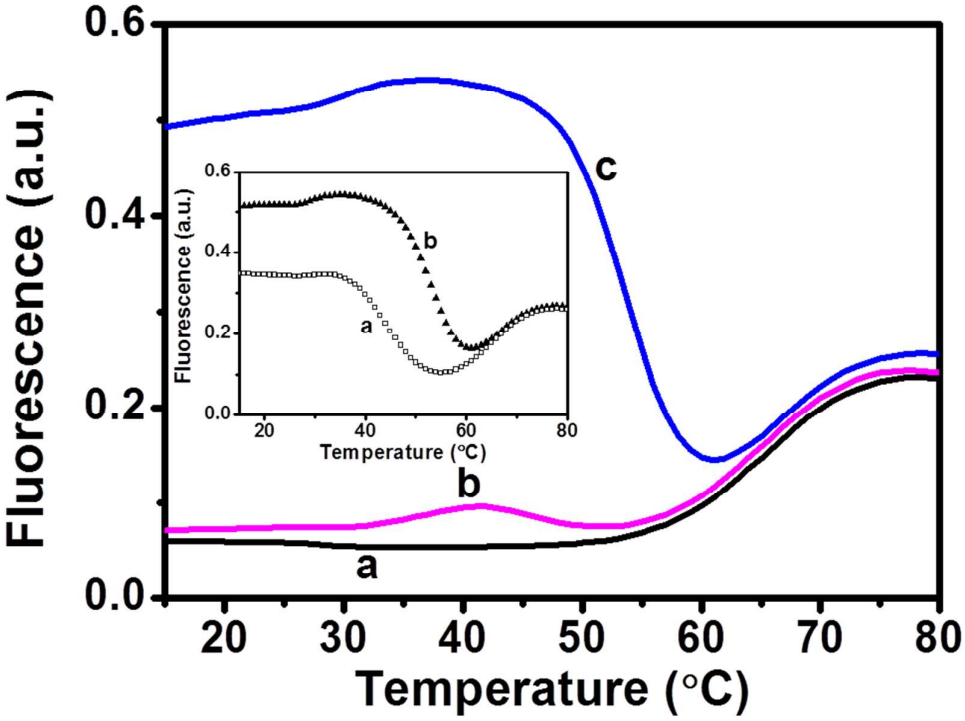


Figure S-1. Thermal denaturation profiles observed as a result of MB1 unfolding in the presence of (a) no target sequence, (b) targets TG and TI, and (c) targets TE and TF. Inset: thermal denaturation profiles of MB1 treated with either target TF (a) or TE (b) alone. Sequences of MB1 and targets are elaborated in Table S-1. The concentrations of MB1 and each target were 0.2 and 1.2 μM, respectively.

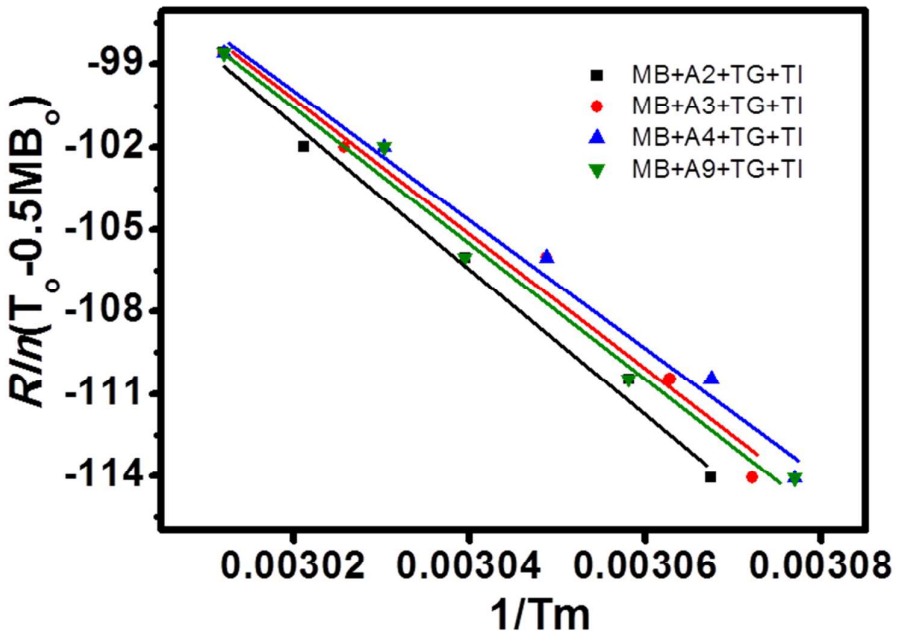


Figure S-2.The relationships between $Rln([T_0] - 0.5[MB_0])$ and reciprocal of melting temperature in connection to varied compositions of space nucleotides.

We sought to harness the 18-nt fragment (in H5) and the 26-nt fragment (in N2) as targets to configure the corresponding hairpin and assistant strand. While we have identified the 18-nt H5-specific fragment, the 8th and 9th nucleotide (T and G) are unfortunately exceptional to the H5N2 strains (e.g., A/duck/Taiwan/DV30-2/2005) employed. This rendered us to perform the shorter (16-nt) candidate fragment instead. Also note that a stable duplex construction between the target and the assistant strand was desired towards a successful operation of the scheme, thereby the longest candidate fragment (i.e., the 26-nt) was exerted for N2-specific sequence to allow a pendent overhang, apart from the segment in hybridization with the loop region, for formation of a long duplex between N2 target and assistant strand. We assigned 8-nt fragment (residing in 3'end of the H5-specific sequence) and 9-nt (5'end in the N2-specific sequence) to hybridize with loop and enable 8- and 16-nts, in H5 and N2, respectively, hybridizing with the assistant strand. This design (Figure S-3), nevertheless, was eventually unsuccessful to function in AND gate format, as the assistant strand/H5 duplex was unfavourably formed due to a short number of base pairs in hybridization.

We therefore turned to render the 16-nt H5-specific sequence in hybridization with loop and employ its neighboring fragment (conserved region of the H5 coding gene, with a deoxyadenylate nucleotide in separate from the specific sequence) to hybridize with the assistant strand. Such the design conferred H5-target sufficient length for hybridization with both the loop portion and the assistant strand, obviating the doubt forming thermodynamically unfavorable duplexes. Uniquely we assigned the deoxyadenylate to serve as the unpaired nucleotide to alleviate the stress tightened in the four-way junction. Once the sequence of molecular beacon was compiled in accordance with the consideration illustrated above, one should note that a deoxythymidylate residing at the 17th nt of H5-target (numbered from 5'end) was predicted to de-hybridize with loop (based on the mfold simulation, Figure S-4). In order to exclude the possibility leaving a single-stranded nucleotide in the loop, an assistant strand was designed to refer the deoxythymidylate as one of the space nucleotide to conclude a molecular beacon in response to a 15-nt H5-specific fragment.

In the N2 aspect, a straight-forwarding approach drove us to assign a comparable length (15-nt, numbered from 5'end) of the N2-specific sequence (26-nt) for the segment in hybridization with loop portion. The remaining 11-nt fragment was thereby used for the duplexes in connection with the assistant strand. In such the circumstance, the 26-nt N2-specific sequence, nevertheless, favorably folded in dimer (simulated by mfold, Figure S-5) in connection to a weak association with the assistant strand. This hampered facile construction of the four-way junction and inhibited the expedition of AND gate processing. A compromising approach to reduce the length (from 15- to 12-nt) in hybridization with loop enabling an extension (from 11- to 14-nt) of target/assistant strand duplex was thereby evolved. In addition, in light of the short duplex existed between N2 target and loop portion,

no space nucleotide was assigned to frank in the N2 side.

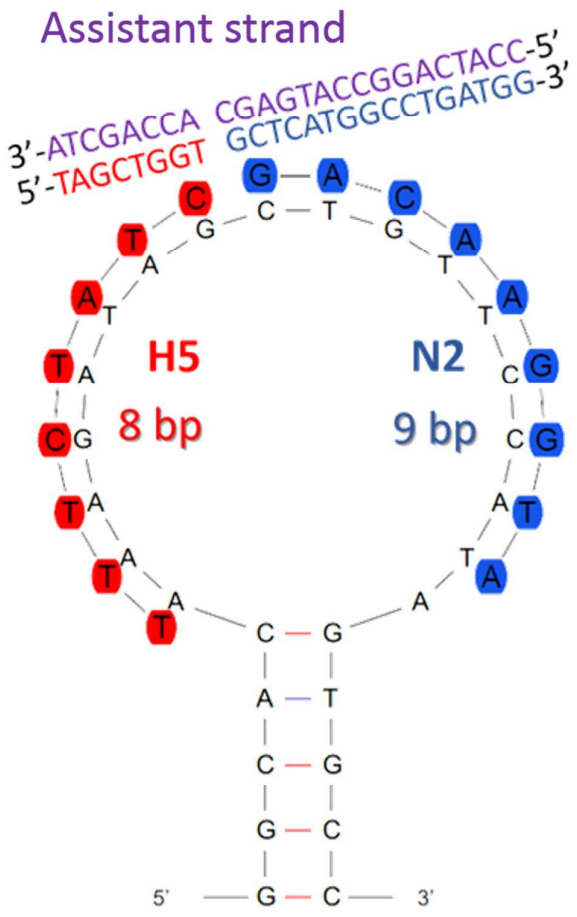


Figure S-3.Envisaged diagram depicting the associate bearing 8-nt fragment (residing in 3'end of the H5-specific sequence) and 9-nt (5'end in the N2-specific sequence) to hybridize with loop, as well as enable 8- and 16-nts, in H5 and N2, respectively, hybridizing to the arming probe.

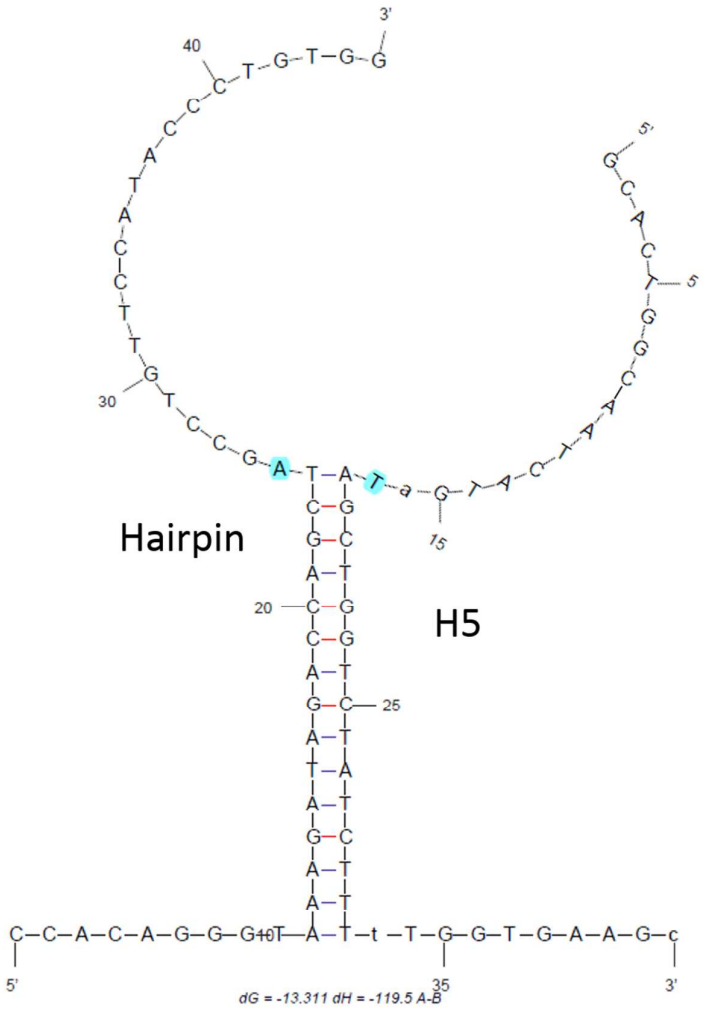


Figure S-4.Mfold simulated result showing that a deoxythymidylate residing at the 17th nt of H5-target (numbered from 5'end) de-hybridized with loop.

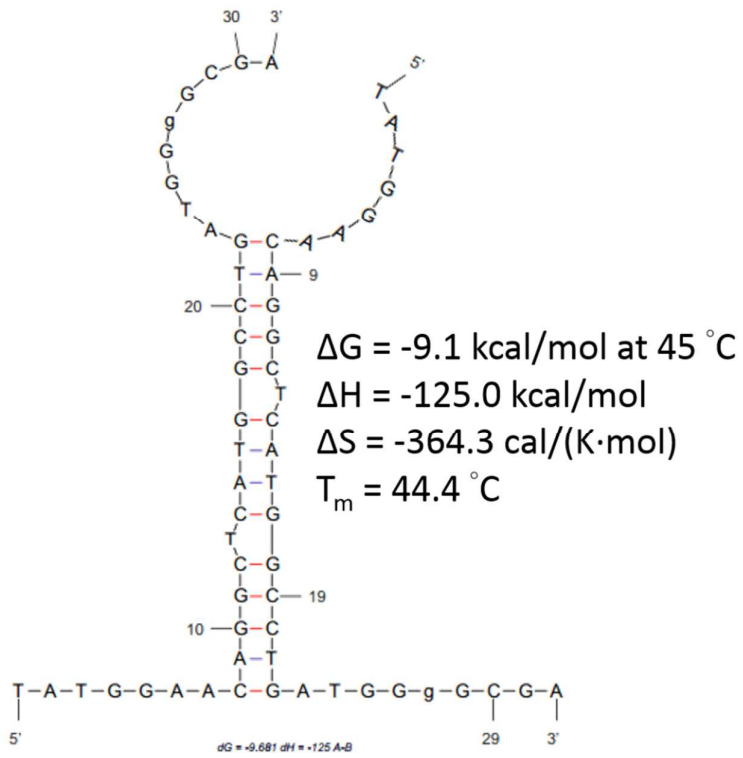


Figure S-5.Mfold simulated result depicting a dimer structure formed by the 26-nt N2-specific sequence.

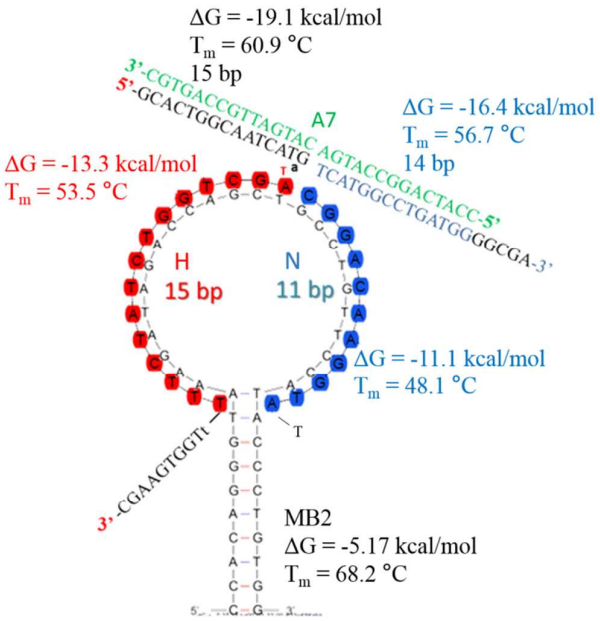


Figure S-6. Schematic configuration of the four-way junction composed of molecular beacon 2 (MB2), the assistant strand (A7), and specific sequences coded in H5 and N2 genes.

Sequence selections of H5- and N2-specific genes

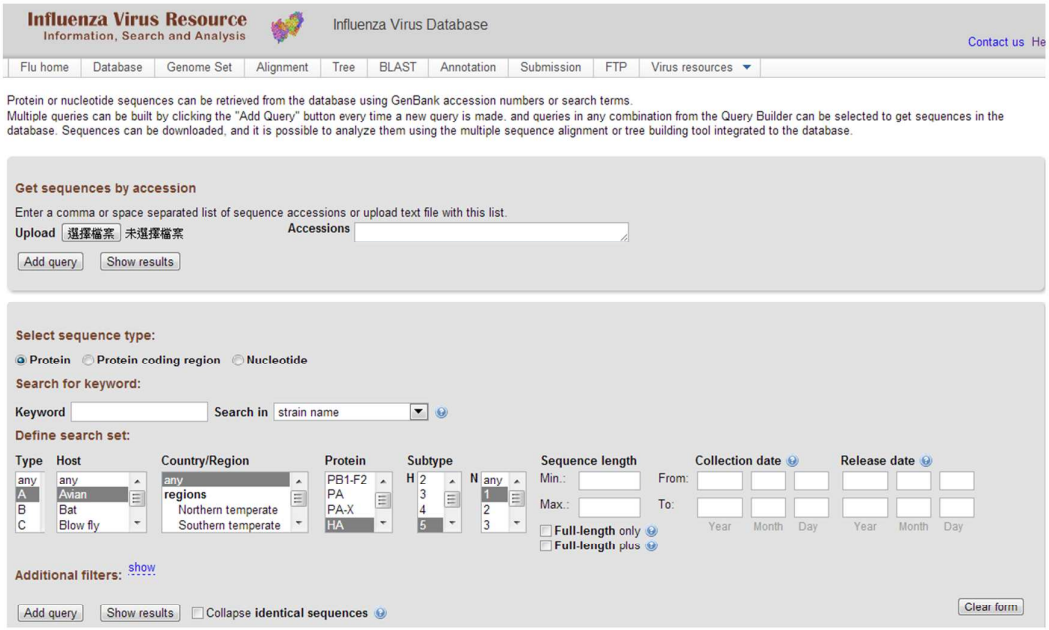
The constituent sequences of MBs were designed based on avian sequences acquired from the Influenza Virus Resource.[2] The downloaded influenza H5 and N2 nucleic sequences were respectively aligned by MUSCLE[3] for identification of contiguous conserved regions in the gene segments. The identified regions thus possessed more than 90% sequence conservation at each base in the whole population of the specified H5N2 subtype. These candidate regions were subsequently aligned to non-H5 and non-N2 subtypes separately with Bowtie[4] to eliminate the promiscuous sequences which can be observed in other subtypes. The selected candidate regions were further analyzed by performing BLAST analysis against Nucleotide collection to assess the sequence specificity of influenza virus.[5] Result snapshots generated through the selection process are detailed below. The resulted H5- and N2-specific sequences are summarized in Table S-2.

Table S-2. H5- and N2-specific sequences identified from the selection processes

Numbered Position	Sequence (5'→3')	Notes
H5		
1613-1630	TCA ACA GTG GCG AGT TCC	18-nt
1650-1665	TAG CTG GTC TAT CTT T	16-nt
N2		
932-948	ATA AAT ATG GCA GAT TA	17-nt
1004-1019	GAT GAT AGC TCT AGC A	16-nt
1202-1221	CAA GTC ATA GTT GAC AAT AA	20-nt
1376-1401	TAT GGA ACA GGC TCA TGG CCT GAT GG	26-nt

Flowchart of sequence selection:

1. Download HA/NA sequences of avian influenza HxN2/H5Nx viruses from Influenza Virus: Resource (http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html).



HA	H1N2	32	NA	H5N1	2702
	H2N2	84		H5N2	472
	H3N2	201		H5N3	90
	H4N2	52		H5N4	4
	H5N2	624		H5N5	5
	H6N2	556		H5N6	1
	H7N2	346		H5N7	8
	H8N2	2		H5N8	7
	H9N2	2465		H5N9	14
	H10N2	12		H5N10	0
	H11N2	68			
	H12N2	3			
	H13N2	14			
	H14N2	0			
	H15N2	1			
	H16N2	0			
	H17N2	0			

2. Align HA/NA sequences of avian influenza H5N2 viruses to reference strains (CY139327 and CY014851 for HA and NA sequence, respectively) using MUSCLE:

For example, the alignment of 642 H5 sequences is illustrated as the following:

```
>CY034681 A/chicken/Pennsylvania/3609/1993 1993// 4 (HA)
GGTCTAAACTATCAAAATGGAAAGAATGGTGATTGCCCTTGCAATAATCAGCATTGTCAAAGGTGACCAAATCTGCATTGGTTATCATGCAAACAATTCAACAGAGCAGGTTGACACAATCATGGAGAAGAATGTAACGGTCACACATGCTCA
>GU052787 A/chicken/Pennsylvania/21525/1983 1983// 4 (HA)
GGTATAATTCATCAAAATGGAAAGAACAGTGATTGCCCTTGCAATAATCAGCGTTGTCAAAGGTGACCAAATCTGCATCGGTTATCATGCAAACAATTCAACAAAGCAAATTGACACAATCATGGAAAAGAATGTGACGGTCACACATGCTCA
>EF607878 A/pheasant/MD/4457/1993 1993// 4 (HA)
------------------------------------------------------------------------------------------------------------------------------------------ATGGAGAAGAATGTGACGGTCACACATGCTCA
>FJ520095 A/northern pintail/California/44221-789/2006 2006/10/05 4 (HA)
--------------------------GTAGTGATTGCCCTCGCAATAATCAGCATTGTCAAAGGTGACCGGATCTGCATTGGTTATCATGCAAACAATTCAACAGAGCAAGTCGATACAATCATGGAGAAGAACGTGACGGTCACACATGCTCA
>AF098537 A/turkey/Ramon/73  1973// 4 (HA)
-----------------------------------------------------------CTTGTCAGAAGTGATCAGATTTGCATTGGTTACCATGCAAACAACTCAACAGAGCAGGTTGACACAATAATGGAGAAGAATGTGACTGTCACACATGCCCA
>AB275425 A/chicken/Ibaraki/8/2005 2005// 4 (HA)
----------------ATGAAAAGAATAGTAATTGCCTTTGCAATAATCAGCATTGTTACAGGTGACCAAATCTGCATTGGTTATCATGCAAACAATTCAACAAAACAAGTTGATACAATCATGGAAAAGAATGTGACGGTCACACATGCTCA
>AB507264 A/chicken/Taiwan/A703-1/2008 2008// 4 (HA)
--TACCATTTATCAAAATGGAAAGAATAGTGATTGCCTTTGCGATAGTCAGCATTGTCACAGGTGACCGAATCTGCATTGGTTATCATGCAAACAATTCAACAAAACAGGTTGACACAATCATGGAAAAGAATGTGACGGTCACACATGCTCA
>L46586 A/chicken/Puebla/8623-607/1994 1994// 4 (HA)
-----------------------------------GCCTTTGCAATGATCAGCATCGTCACAGGTGACCAAATCTGCATTGGTTATCATGCAAACAATTCAACAAAACAGGTTGACACAATCATGGAGAAGAATGTGACGGTCACACATGCTCA
>CY017403 A/poultry/Italy/330/1997 1997// 4 (HA)
-------TCTGTCAAAATGGAGAAAATAGTGCTTCTTCTTGCAATAGTTAGTCTTGTTAAAAGTGACCAGATTTGCATCGGTTACCATGCAAACAACTCGACAGAGCAGGTTGACACAATAATGGAAAAGAATGTTACTGTAACACATGCCCA
>CY133809 A/gadwall/Missouri/10MO0280/2010 2010/11/16 4 (HA)
GGTCCAAACTATAAAAATGGAAAGAATAGTGATTGCCCTCGCAATAATCAGCATTGTCAAAAGTGACCAAATCTGCATTGGTTATCATGCAAACAATTCAACAGAGCAGGTTGATACAATCATGGAGAAGAATGTGACGGTCACACATGCCCA
>GU052612 A/chicken/Mexico/22184/1998 1998// 4 (HA)
------------------------------------------------------------------------------------------------------------------------------------------GAGAAGAATGTGACGGTCACACATGCTCA
>GQ923373 A/waterfowl/Colorado/476466-2/2007 2007// 4 (HA)
GGTTCAAACCATGAAAATGGAAAGAATAGTGATTGCCCTCGCAATAATCAGCATTGTCAAAGGTGACCAAATTTGCATTGGTTACCATGCAAACAATTCAACAGAGCAGGTTGATACAATCATGGAAAAGAATGTGACGGTCACACATGCTCA
>AY296069 A/Avian/NY/31588-2/2000 2000// 4 (HA)
-------------------------------------------------------------------------------ATCTGCATTGGTTATCATGCAAACAATTCAACAGAACAGGTCGACACAATCATGGAGAAGAATGTGACGGTCACACATGCTCA
>AF164656 A/turkey/Virginia/40018/84 1984// 4 (HA)
-------------------------------------------------------------------GACCAAATCTGCATCGGTTATCATGCAAACAATTCAACAAAGCAAATTGACACAATCATGGAAAAGAATGTGACGGTCACACATGCTCA
>CY107847 A/chicken/Pennsylvania/1/1983 1983// 4 (HA)
----------------ATGGAAAGAACAGTGATTGCCCTTGCAATAATCAGCGTTGTCAAAGGTGACCAAATCTGCATCGGTTATCATGCAAACAATTCAACAAAGCAAATTGACACAATCATGGAAAAGAATGTGACGGTCACACATGCTCA
>U79450 A/mallard/OH/345/1988 1988// 4 (HA)
-------------------------------------------------------------------GACCAAATCTGCATTGGTTATCATGCAAACAATTCAACAGAACAGGTTGACACAATCATGGAGAAGAATGTGACGGTCACACATGCTCA
>GQ117169 A/duck/New York/466787/2006 2006// 4 (HA)
GGTCCAAACTATGAAAATGGAAAGAATAGTGATTGCCCTCGCAATAATCAGCATTGTCAAAGGTGACCAAATTTGCATTGGTTACCATGCAAACAATTCAACAGAGCAGGTTGATACAATCATGGAAAAGAATGTGACGGTCACACATGCTCA
>CY138312 A/blue-winged teal/New Brunswick/00288/2010 2010/09/14 4 (HA)
GGTCCAAACTATGAAAATGGAAAGAATAGTGATTGCCCTCGCAATAATCAGCATTGTCAAAGGTGACCAAATTTGCATTGGTTACCATGCAAACAATTCAACAGAGCAGGTTGATACAATCATGGAAAAGAATGTGACGGTCACACATGCTCA
>GU050144 A/gadwall/California/490899/2007 2007// 4 (HA)
GGTCCAAACTATAAAAATGGAAAAAATAGTGATTGCCATCGCAATAATCAGCATTGTCAAAGGTGACCGGATCTGCATTGGTTATCATGCAAACAATTCAACAGAGCAAGTCGATACAATCATGGAGAAGAACGTGACGGTCACACATGCCCA
>FJ648285 A/mallard/Bavaria/1/2007 2007/09/15 4 (HA)
----------------ATGGAGAAAATAGTACTTCTTTTTGCAGTAGTCAGTCTTGTCCAAAGTGACCAGATTTGCATTGGTTACCATGCAAACAACTCGACAGAGCAGGTTGACACAATAATGGAAAAGAATGTTACTGTCACGCATGCCCA
>CY049904 A/northern pintail/Wisconsin/462764/2006 2006// 4 (HA)
```

3. Calculate nucleotide frequency at each position: If the frequency>0.9, the position is regarded as conserved in this subtype and shown as capitalized letter. Therefore, several contiguous conserved regions can be selected (as highlighted in red rectangles).

```
0 G:201 A:1 T:1 -:421
1 G:201 A:1 T:1 -:421
2 T:214 -:410
3 C:188 A:23  T:8 -:405
4 C:157 T:63  A:1 -:403
5 A:202 C:14  G:4 T:1 -:403
6 A:219 T:2 C:1 -:402
7 A:168 T:53  G:7 -:396
8 C:203 T:17  A:9 -:395
9 T:207 C:22  A:1 -:394
10  A:194 G:37  -:393
11  T:230 C:2 -:392
12  C:80  G:79  A:69  T:3 -:393
13  A:237 -:387
14  A:237 G:3 -:384
15  A:240 G:2 -:382
16  A:426 -:198
17  T:426 -:198
18  G:427 -:197
19  G:405 A:22  -:197
20  A:423 G:2 C:1 -:198
21  A:270 G:156 -:198
22  A:422 G:4 -:198
23  G:242 A:184 -:198
24  A:421 G:4 C:1 -:198
25  A:396 G:37  C:1 -:190
26  T:414 C:21  A:1 -:188
27  A:434 G:3 T:1 -:186
28  G:436 A:4 -:184
29  T:440 -:184
30  G:328 A:111 C:1 T:1 -:183
```

```
>Frequency_Seq
GGTccAAacTaT-AAAATGGAaAgAATAGTgaTTgcccTtGCAATAaTCAGcaTtGTcAAAgGTGACCaaATcTGCATTGGTTAHCATGCAAACAAtTCa
ACAgAgCAgGTtGAcACAATcATGGAaAAGAAtGTgACgGTCACaCATGCtCAgGAcATACTgGAaAAagagCACAAtGGgAaaCTcTGCAGtCTtAAaG
GAGTgAagCCcCTCATtcTGAaGGATTGcAGtGTAGCTGGaTGGCTtCTtGGAAAcCCaATGTGTGAtGAATTCCTgAATGTaCCgGAATGGTCaTACAT
-GTgGAaAAaGAcAatCCAgtCAATGGCCTgTGcTAtCCaGGagAcTTCaacGAtTAtGAAGAaCTGAAgCAttTaaTGAGcAGcACAAAcCATTTTGAG
AAAATTCAgATaaTcCCtAGgAGtTCTTGGtCCAAtCATgATGCCTCATCAGGaGTGAGcTC-GCATGcCCaTAcaaTGGtAGgTCtTCcTTTTTcaGgA
ATGTaGTgTGGtTgATCAAgAAGaAtAATGcgTACCcaACAaTAAAgAGGAccTAcAacAAcACcAAtgtAGAAGAcCTTtTaaTA-TaTGGGGaATtCA
cCAcCCTAATGATGCaGCtGAaCAaAcaAAaCTCTAcCAgAACtcgAaCACtTAtGTgTCtGT-GGaACATCAACACTGAAtCAgAGaTCaaTcCCAgAA
ATAGCcACcAGaCCCaAaGTgAAcGGaCAaAGtGGAAGaATGGAaTTtTTcTGGACAATacTaAaGcCgAAcGATgCaATCa-cTTtGAgAGtAaTGGgA
AtTTTATaGCTCCtGAATATGCaTACAAgATTgTcAAgAAaGGagA-TCAGCAATCATGAaAAGTGaatTGGAgTAtGGTAACTGtgAcaCcAAaTGtCA
gACcCCA-TgGGTGCtATAAAtTCcAGtaTGCC-TTcCACAAtgT-CAtCCtcTtACCATTGGgGAgTGcCCCAagTAtGTcAAaTCggA-AaAcTgGTc
CTTGCaACaGGaCTaAGaAAcGTaCCcCAAAgAgAAAcAAGAGGccTATTTGGaGCaATAGCaGGaTTcATAGAAGGAGGaTGGCAAGGaATGGT-GAtG
G-TGGTAtGGaTAcCAtCATAGcAAtGAGCAgGGaAGTGGaTAtGCTGCAGACAAAGAaTCtACcCAgAAaGCAATcGATGGgATCACcAATAAaGT-AA
CTCAATCATTGACAAAATGAACACtCAaTTcGAaGCcGTTGGgAAaGAATTcAAcAACcTaGAAAGgAGaATAGAaAATTTgAAtAAGAAaATGGAAGAt
GGgTTttTaGATGTaTGGACTTAcAATGCaGAACTTCT-GTgCTCATGGAAAAtGAaAGAACtcTgGAttTcCATGAtTCAAATGTCAAGAACcTaTAcG
AtAAGGTcCGACTcCAGCTgAGaGAcAATGCAAAaGAatTgGGcAAtGG-TGcTTtGAaTTCTAcCACAAgTGTGAcaATGAATGcATGGAAAGTGTgAG
AAAtGGAACgTATgAcTAtCCgCAaTAtTCAGAAGAatCAAGAcTgAAcAGaGAGGAAATAgacGGAGTcAAATTgGAATCAATgGGcACcTAtCAgATA
cTaTCAATcTAcTCAACAGTGGCGAGTTCCcTAGCACTGGCAATCATGaTAGCTGGTCTATCTTTtTGGATGTGcTCCAATGGaTCATTGCAgTGCAGaA
TTTGCATcTAgaaTTGTGAGTTCAGATTaTAaTTAAAAACACC
```

4. To identify HA/NA-specific sequence: the contiguous conserved regions were aligned to non-H5/N2 sequences using Bowtie and discarded the ones hitting to non-H5/N2 sequences.

```
3_Alignment_11.sam
Candidate_1 H10N2:1 H11N2:77  H12N2:3 H13N2:12  H1N2:22 H2N2:53 H3N2:117  H4N2:27 H6N2:547  H7N2:911  H8N2:5  H9N2:5208
Candidate_2 H10N2:6 H11N2:8 H12N2:3 H13N2:22  H15N2:1 H1N2:27 H2N2:99 H3N2:150  H4N2:72 H6N2:732  H7N2:32 H8N2:2  H9N2:5470
Candidate_3 H11N2:10  H1N2:8  H2N2:37 H3N2:361  H4N2:10 H6N2:2  H7N2:242  H8N2:2  H9N2:157
Candidate_4 H10N2:1 H11N2:93  H12N2:3 H13N2:16  H15N2:1 H1N2:39 H2N2:149  H3N2:6  H4N2:20 H6N2:813  H7N2:572  H8N2:4  H9N2:3577
Candidate_5 H10N2:8 H11N2:61  H12N2:4 H13N2:20  H15N2:2 H1N2:43 H2N2:142  H3N2:138  H4N2:101  H6N2:856  H7N2:471  H8N2:2  H9N2:132
Candidate_6 H10N2:37  H11N2:90  H13N2:12  H15N2:1 H1N2:35 H2N2:163  H3N2:78 H4N2:42 H6N2:117  H7N2:332  H8N2:2  H9N2:1975
Candidate_7 H10N2:13  H11N2:70  H12N2:8 H13N2:10  H15N2:3 H1N2:28 H2N2:68 H3N2:231  H4N2:62 H6N2:322  H7N2:104  H8N2:2  H9N2:1647
3_Alignment_13.sam
Candidate_1 H15N2:1 H1N2:2  H3N2:148  H4N2:11 H6N2:127  H8N2:2
Candidate_2 H1N2:2  H2N2:161  H3N2:7  H7N2:2  H8N2:2  H9N2:12
3_Alignment_14.sam
Candidate_1 H2N2:55 H3N2:53 H6N2:146  H7N2:5  H9N2:23
Candidate_2 H11N2:67  H13N2:14  H2N2:79 H4N2:20 H6N2:293  H9N2:626
3_Alignment_15.sam
Candidate_1 H1N2:2  H2N2:30 H6N2:152  H8N2:2
3_Alignment_16.sam
_
3_Alignment_17.sam
Candidate_1 H2N2:8
Candidate_2 H2N2:70
3_Alignment_18.sam
_
3_Alignment_26.sam
Candidate_1 H1N2:3  H6N2:117
```

Specific to H5

REFERENCES

(1) Hoffmann, E.; Stech, J.; Guan, Y.; Webster, R. G.; Perez, D. R. *Arch. Virol.* **2001**, *146*, 2275-2289.

(2) Bao, Y. M.; Bolotov, P.; Dernovoy, D.; Kiryutin, B.; Zaslavsky, L.; Tatusova, T.; Ostell, J.; Lipman, D. *J. Virol.* **2008**, *82*, 596-601.

(3) Edgar, R. C. *Nucleic Acids Res.***2004**, *32*, 1792-1797.

(4) Langmead, B.; Trapnell, C.; Pop, M.; Salzberg, S. L. *Genome Biol.***2009**, *10*.

(5) Johnson, M.; Zaretskaya, I.; Raytselis, Y.; Merezhuk, Y.; McGinnis, S.; Madden, T. L. *Nucleic Acids Res.***2008**, *36*, W5-W9.