Surveillance of genetic drift in Influenza A virus with base-specific cleavage and MALDI-TOF mass spectrometry

Gao, Xiang

2013

Gao, X. (2013). Surveillance of genetic drift in Influenza A virus with base-specific cleavage and MALDI-TOF mass spectrometry. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/53730>

<https://doi.org/10.32657/10356/53730>

NANYANG
TECHNOLOGICAL
UNIVERSITY

**Surveillance of genetic drift in Influenza A virus with
base-specific cleavage and MALDI-TOF mass spectrometry**

Gao Xiang

SCHOOL OF BIOLOGICAL SCIENCES

A thesis submitted to the Nanyang Technological University in partial
fulfilment of the requirement for the degree of

Doctor of Philosophy (Ph.D)

2013

Acknowledgements

I would like to express my deepest respect and most sincere gratitude to my supervisor, Dr. Kai Tang, for his valuable guidance and encouragement. All my research work is under his inspirations. He also imparts his knowledge and experience to me and gives me valuable suggestions for the experiments. I would like to give thanks to A.Prof. Richard Sugrue and Dr. Tan Boon Huan, for their kindly guidance in virology and influenza study. My deepest thanks also go to Dr. Mu Yuguang, Dr. Li Jinming and Dr. Chen Xin, for their help in statistical data analysis and Mr. Yong Chee Tan for his contribution in compound matrix analysis.

I would like to convey my gratitude to my colleagues in NTU: Amm, Xiao Jie, Feng Tao, Dong Xueming, Nantika, Han Nanyu, Zheng Jie, who offered me great help and assistance during my Ph.D period. Special thanks go to my best friend Li Fupeng, who gives me irreplaceable suggestions and help through the past few years.

Last but not least, I wish to thank my parents and my girlfriend Ms. Zhang Jie for their love and support.

Abbreviations

3-HPA	3-hydroxypicolinic acid
ACN	acetonitrile
ATT	6-aza-2-thiothymine
ddNTP	dideoxynucleotide triphosphate
ddH ₂ O	double distilled water
EB	ethidium bromide
ESI	electrospray ionization
FFPE	formalin-fixed paraffin-embedded
HA	hemagglutinin
HPV	human papillomavirus
MALDI	matrix-assisted laser desorption/ionization
MLEE	multilocus enzyme electrophoresis
MLST	multilocus sequence typing
MS	mass spectrometry
NA	neuraminidase
PCR	polymerase chain reaction
PFGE	pulsed-field gel electrophoresis
PMF	peptide mass fingerprinting
RDBH	reverse dot blot hybridization
RFLP	restriction fragment length polymorphisms
SIV	swine influenza virus

S/N	signal to noise ratio
SABER	single allele base extension reaction
SAP	shrimp alkaline phosphatase
SNP	single-nucleotide polymorphism
TFA	trifluoroacetic acid
TIGER	Triangulation Identification for the Genetic Evaluation of Risks
TOF	time of flight

TABLE OF CONTENTS

Summary	10
Chapter One : Introduction.....	12
1.1 Nucleic Acid Sequences and Genes.....	12
1.2 SNP	13
1.3 Genotype and Genotyping	14
1.4 Analysis of Nucleic Acid with MALDI-MS	15
1.4.1 Brief History of MALDI	15
1.4.2 Ionization Mechanism	16
1.4.3 Matrix for Nucleic Acid Analysis	18
1.4.4 Sample Preparation	19
1.4.5 Mass Spectrometry Instrumentation	21
1.5 Primer Extension Assay.....	22
1.5.1 SNP Genotyping.....	22
1.5.2 Genome-Wide Association Study.....	27
1.5.3 Gene Expression Analysis	29
1.5.4 Non-Invasive Prenatal Diagnosis	30
1.6 Base-Specific Cleavage Assay	31
1.6.1 SNP Discovery	31
1.6.2 Identification of tRNA	36
1.6.3 RNA Modifications Analysis	36
1.6.4 DNA Methylation Analysis	37
1.6.5 MLST for Bacterial Studies	38
1.6.6 Pathogenic Viruses Study	40
1.7 Influenza A Virus	41
1.8 Research Objectives	46
1.8.1 Comparative Sequencing of H1N1 Influenza Virus.....	46
1.8.2 Surveillance of Genetic Drift	47

Chapter Two Materials and Methods.....	48
2.1 Materials and Reagents.....	48
2.1.1 Virus	48
2.1.2 RT-PCR	48
2.1.3 PCR	48
2.1.4 DNA Analysis.....	48
2.1.5 SAP Treatment	49
2.1.6 <i>in vitro</i> Transcription	49
2.1.7 RNaseA Digestion.....	50
2.1.8 Resin.....	50
2.1.9 MALDI.....	50
2.2 Methods	50
2.2.1 Viral cDNA Preparation	50
2.2.2 Base-Specific Cleavage Assay	51
2.2.3 MALDI-TOF MS Analysis	53
2.2.4 <i>in silico</i> Digestion.....	53
Chapter Three : Analysis of Two Genes from Two Viral Strains	56
3.1 Analysis of HA Genes from WSN &PR8	56
3.1.1 Fragment One.....	58
3.1.2 Fragment Two	68
3.1.3 Fragment Three	74
3.1.4 Fragment Four.....	80
3.1.5 Coverage of HA gene	85
3.2 Analysis of HA genes from WSN/PR8 mixture.....	86
3.3 Analysis of NA genes from WSN & PR8	90
3.3.1 Fragment One.....	92
3.3.2 Fragment Two	97
3.3.3 Fragment Three	102
Chapter Four : Surveillance of genetic drift.....	108
4.1 Validation of Known Sequence Changes	108

4.2 Monitoring of Unknown Sequence Changes	114
4.3 Surveillance of genetic drift.....	117
Chapter Five : Data Analysis & Optimization.....	123
5.1 Evaluation of the MS data	123
5.2 Cyclic Phosphate.....	128
5.3 Optimization of Matrix	136
Chapter Six: Conclusion.....	139
References	142

LIST OF FIGURES

Figure 1. A Schematic of MALDI-TOF MS Machine.	17
Figure 2. Primer Extension vs Pinpoint Assay.	25
Figure 3. Schematic diagram of the base-specific cleavage assay.	35
Figure 4. schematic structure of Influenza virus.	43
Figure 5. Stucture of amantadine and rimantadine.....	45
Figure 6. Stucture of oseltamivir	46
Figure 7. Mongo Oligo Mass Calculator v2.06.	55
Figure 8. Four pairs of primers for HA gene.....	56
Figure 9. PCR products on agarose gel..	58
Figure 10. WSN HA F1 +dC.....	63
Figure 11. WSN HA F1 +dU.....	64
Figure 12. WSN HA F1 -dC.....	65
Figure 13. WSN HA F1 -dU.....	66
Figure 14. WSN HA F2 +dC.....	69
Figure 15. WSN HA F2 +dU.....	70
Figure 16. WSN HA F2 -dC.....	71
Figure 17. WSN HA F2 -dU.....	72
Figure 18. WSN HA F3 +dC.....	75
Figure 19. WSN HA F3 +dU.....	76
Figure 20. WSN HA F3 -dC.....	77
Figure 21. WSN HA F3 -dU.....	78
Figure 22. WSN HA F4 +dC.....	81
Figure 23. WSN HA F4 +dU.....	82
Figure 24. WSN HA F4 -dC.....	83
Figure 25. WSN HA F4 -dU.....	84
Figure 26. Venn Diagram of the amount of peaks in HA fragment 3 positive dU reaction between WSN and PR8	87
Figure 27. WSN : PR8 = 5:1	88
Figure 28. WSN : PR8 = 1:3.	89
Figure 29. Three pairs of primers for NA gene	90
Figure 30. WSN NA F1 +dC.....	93
Figure 31. WSN NA F1 +dU.....	94
Figure 32. WSN NA F1 -dC.....	95
Figure 33. WSN NA F1 -dU.....	96
Figure 34. WSN NA F2 +dC.....	98
Figure 35. WSN NA F2 +dU.....	99
Figure 36. WSN NA F2 -dC.....	100
Figure 37. WSN NA F2 -dU.....	101
Figure 38. WSN NA F3 +dC.....	103
Figure 39. WSN NA F3 +dU.....	104
Figure 40. WSN NA F3 -dC.....	105

Figure 41. WSN NA F3 -dU.....	106
Figure 42. +dC 2925Da vs 2965Da.....	111
Figure 43. +dU 3637 Da.....	112
Figure 44. +dU 2650 Da.....	113
Figure 45. Output of automatic analysis.....	115
Figure 46. WSN-(T5A) HA F3 -dC.	119
Figure 47. WSN-(T5A) HA F4 +dC.....	121
Figure 48. WSN-(T5A) HA F4 +dU..	122
Figure 49. Spectrum Quality vs Match.....	123
Figure 50. High Quality +well Match	125
Figure 51. High Quality + poor Match.....	126
Figure 52. Low Quality + well Match --- Charge competition	127
Figure 53. 2',3'-cyclic phosphodiester intermediate	128
Figure 54. Cyclic phosphate dC 1543+1559.	131
Figure 55. Cyclic phosphate dC 1872+1888.....	132
Figure 56. Cyclic phosphate dU 2217	133
Figure 57. Cyclic phosphate dU 2602	134
Figure 58. Cyclic phosphate dU 1968	135
Figure 59. Spectra from 3-HPA,ATT and Compound Matrix	137
Figure 60. Microscopic view of matrix crystals	138

LIST OF TABLES

Table 1. Primers for HA genes.	57
Table 2. <i>in silico</i> and experimental peak numbers of WSN HA gene	86
Table 3. Primers for NA genes.	91
Table 4. <i>in silico</i> and experimental peak numbers of WSN NA gene	107
Table 5. <i>in silico</i> and experimental peak numbers of PR8 NA gene	107
Table 6. Comparison of sequencing result with the NCBI database.	109
Table 7. Result of automatic analysis.....	116
Table 8. MW of dNTP and rNTP used in the assay	118

Summary

With the discovery of several matrices which enable the ionization of DNA and RNA, MALDI mass spectrometry has become a powerful platform for the study of nucleic acid sequence changes, (e.g. mutations, single nucleotide polymorphisms (SNPs), insertion/deletion, alternative splicing, etc.), amount changes (e.g. copy number variation, gene expression, allele expression, etc.), as well as modifications (e.g. methylation of genomic DNA, post transcriptional modification of tRNAs and rRNAs). Two major strategies have been employed to characterize these changes. Primer extension reactions are designed for genotyping of known polymorphic sites and determining the levels of gene or allele expressions. Base-specific cleavage reactions are used for discovery of unknown polymorphisms and characterization of modifications. These two assays usually generate nucleic acid fragments less than 30 bases long, which is the ideal mass range for MALDI mass spectrometry.

Treatments of influenza infection are most effective if administered within 36 to 48 hours after the onset of symptoms. To develop a rapid and inexpensive diagnostic method, base-specific cleavage followed by MALDI-TOF mass spectrometry analysis has been used for high throughput and high resolution identification of all species of influenza viruses. A signature region of 300 to 800 base pairs in length from the viral genome is transcribed and then digested with RNase A to produce a strain-specific set of mass peaks, which serve as a molecular weight fingerprint. This approach has great potential in the surveillance of global spread and emergence of novel viral genotypes especially from antigenic drift.

In this project, the approach is evaluated using Influenza A/WSN/1933 and A/PR/8/1934 viral strains. The molecular weight fingerprints obtained from mass spectrometric data cover more than 90% of the genomic sequence in query, and in perfect agreement with DNA sequencing result, suggesting that the rapid identification method is feasible and accurate.

Chapter One. Introduction

1.1 Nucleic Acid Sequences and Genes

Nucleic acids are biological molecules which are polymers of nucleotides. A nucleotide consists of a base, a pentose sugar and a phosphate group. Nucleic acids are divided into two groups - DNA (deoxyribonucleic acids) and RNA (ribonucleic acid), according to the type of the pentose sugar that builds up the nucleotide – a D-ribose or a D-deoxyribose.

DNA consists of two poly-nucleotide chains, with backbones made of pentose sugars and phosphate groups. There are four types of bases that can be attached to each pentose sugar – cytosine (C), guanine (G), adenine (A) and thymine (T). The two strands of a DNA molecule run in anti-parallel via specific pairing of bases: C is paired with G and A is paired with T. The main role of DNA molecules is the long-term storage of genetic information which is encoded by the sequence of the bases along the DNA backbone.

RNA is also made up of a chain of nucleotides, but the type of bases in RNA is slightly different from that in DNA – thymine is replaced by uracil (U). Some RNA can carry genetic information in the sequence of nucleotides like DNA (1), while some other RNA can have enzymatic activity just like proteins.

The sequence of nucleotides makes one DNA or RNA molecule different from another. It carries genetic information thus is critical for all living organisms. The unit of heredity that resides on a stretch of nucleic acid sequence is called a “gene”. Genes carry the basic instructions which encode for the synthesis and assemblies of all bio-molecules, thus decide

the behavior of one cell or organism (2-4).

1.2 SNP

A number of variations could be found when DNA sequences from different individuals are compared. A variation found in more than 1% of the population is defined as a polymorphism (5). In 1978, the discovery of “restriction fragment length polymorphisms” (RFLP) was reported (6) and the advancement of the RFLP technique opened the door for the study of polymorphisms. Homologous DNA molecules are digested by restriction enzymes and the resulting restriction fragments appear different in length after gel electrophoresis separation. This difference in the length of restriction fragments could happen as a result of a single nucleotide polymorphism (SNP), which could either introduce a new restriction enzyme recognition site or abolish an existing one (5,7). These single nucleotide polymorphisms are of the highest abundance in all types of variations and a lot of related databases are available via internet (8-22). The frequency of SNPs in the human genome is thought to be at least one per one thousand base-pairs (23-27). These SNPs are unevenly distributed across the genome, with only a small portion located in the coding region (28-30). SNPs in coding regions can alter the structure and the function of the protein, which might result in the occurrence of a certain disease or a different response to a certain drug, while other abundant SNPs in non-coding regions are also widely used as important biomarkers in genetic and genomic studies (31).

1.3 Genotype and Genotyping

Genotype is the genetic constitution of a cell, an individual or an organism. Genotyping is the process of determining the genes or genotypes of an individual by examining its DNA sequence with a molecular biological assay. Sanger sequencing, also well-known as “dideoxy sequencing” was first devised in the 1970s by Frederick Sanger and has become the standard technique for the determination of the sequence of DNA molecules (32-35). This sequencing method uses dideoxynucleotide triphosphates (ddNTPs) as chain terminators in an enzymatic DNA synthesis procedure. Sanger sequencing method has been the “gold standard” in the identification of both known and unknown sequence specific nucleotide variations for many years, against which new emerging technologies need to be judged. However, as the Sanger sequencing requires radioactive or fluorescent reporters, and uses separation methods like gel or capillary electrophoresis which generates massive quantity of reads to be analyzed later, the whole process is always rather time-consuming, laborious, and costly. For this reason, researchers are constantly searching for new genotyping technologies (7,36,37), which can meet as many as possible the following requirements:

1. the new assay should be cost efficient.
2. the assay should be of enough sensitivity and reproducibility in case that even tiny amount of DNA samples could still be detected.
3. the assay should be able to detect the relative frequency of individual alleles mixed in pooled samples.
4. the assay should be able to achieve a high level of throughput and automation.

5. the analysis of the result should be simple and automated.

It is not easy to find a single perfect genotyping method which can meet all those different requirements. When a new method is being evaluated, the level of throughput and the efficiency in the time and cost are normally of the most importance (35).

1.4 Analysis of Nucleic Acid with MALDI-MS

1.4.1 Brief History of MALDI

The method of matrix-assisted laser desorption ionization (MALDI) was created by Franz Hillenkamp, Michael Karas and their colleagues in 1985(38). In their research, amino acids and dipeptides were used for ionization. With a pulsed frequency-quadrupled Nd-YAG laser (266 nm), some amino acids and dipeptides (aromatic AAs) show absorbance and the others not (aliphatic AAs). However, the non-absorbing aliphatic amino acids, such as alanine, could be ionized more easily if mixed with absorbing matrix (the amino acid tryptophan). The mechanism for this enhanced ion yield of nonabsorbing molecules in an absorbing matrix was studied (39) and a number of different matrices were found.

In 1987, Koichi Tanaka and his co-workers from Shimadzu Corp. described the “ultra fine metal plus liquid matrix method” for sample preparation, which combined cobalt powder of about 300Å in glycerol with a nitrogen laser (337nm) for ionization. With this laser and matrix combination, the yield of high molecular weight ions (up to 34 kDa protein) was significantly increased and the production time was effectively prolonged(40). Tanaka demonstrated that, a protein could be ionized with a proper combination of laser wavelength and matrix, for which part of the 2002 Nobel Prize in Chemistry was awarded(41).

By the end of 1980s, MALDI-TOF MS was further developed by Hillenkamp and Karas and was able to ionize proteins with molecular weight of 67 kDa(42) and even 100 kDa range(43) using an aqueous solution of nicotinic acid as the matrix and a 266nm laser. Later the application of the frequency-tripled output of a Nd-YAG laser (355nm) and the cinnamic acid derivatives ferulic, caffeic, and sinapinic acids as matrices further improved the MALDI technology(44).

Paralleled with the development of MALDI-TOF technology, another “soft ionization” technique--electrospray ionization (ESI) MS was developed by Fenn(45) and these two methods made it possible for the mass spectrometric analysis of large biomolecules-- proteins and nucleic acids. Nowadays, both MALDI-TOF and ESI-MS are widely utilized in proteomics and genomics research.

1.4.2 Ionization Mechanism

In order to analyze a protein or a nucleic acid with MALDI-TOF MS (Fig. 1), the analyte need to be embedded in a crystalline structure of a certain kind of matrix. With energy from the excitation laser, the co-crystal of the analyte and the matrix is ionized to gaseous phase. The mechanistic principles of this laser induced ionization have been discussed for years(46-53), and it is commonly agreed that the matrix serves as the “launching material” as it absorbs the laser energy and becomes ionized by this event. Then, part of the charge (usually one proton) is transferred from the ionized matrix to the analyte molecules, resulting in the “soft ionization” of the analyte as intact molecules. Both the matrix and the analyte ions are accelerated in an electric field of around 20 kV. Ions with different mass-to-charge ratio

will finally separate after travelling through a field-free drift region. The time of flight (TOF) of the ions could be measured when they reach the detector and this TOF is linearly correlated with the mass-to-charge ratio. As most of the ions generated in the MALDI process are singly charged, the molecular weight information of the analyte could easily be represented in the output spectrum.

Figure 1. A Schematic of MALDI-TOF MS machine. The figure was adopted from a training booklet from Applied Biosystem.

1.4.3 Matrix for Nucleic Acid Analysis

The analysis of oligonucleotides using MALDI-TOF MS was first reported in 1990 (54-56). In the beginning, only combinations of matrix and laser wavelength that worked for protein and peptide analyses were tested, but researchers soon noticed the problems of extensive ion fragmentation and adduction of metal cations to the analyte molecule which resulted in much lower quality of oligonucleotide mass spectra compared to those from proteins and peptides. It was obvious that nucleic acids behave differently from proteins and peptides under MALDI, thus new MALDI conditions, especially new matrices, were required for better ionization.

In 1992, Tang et al. reported the use of a mixture of 3-methylsalicylic acid and 3-hydroxy-4-methoxybenzaldehyde as matrix resulted in the detection of oligonucleotides up to a 34-mer (57), and Nordhoff et al. reported the use of succinic acid, urea and nicotinic acid as matrices in combination with an infra red (IR) laser at wavelength 2.94 μm which led to the detection of tRNAs and 5S rRNA from *Escherichia coli* (58). In 1994, Wu et al. reported the use of 3-hydroxypicolinic acid (3-HPA) as matrix to analyze single-stranded DNA oligomers up to 67 nucleotides in length (59). Although other compounds such as 2,4,6-trihydroxyacetophenone(60), 2,3,4-trihydroxyacetophenone(61), picolinic acid(62,63), 3-aminopicolinic acid(64), 6-aza-2-thiothymine (ATT) (65), 5-methoxysalicylic acid (5-MSA) (66), quinaldic acid (QA) (67), pyrazinecarboxylic acid (PCA) (68), 3-hydroxycoumarin (3-HC) (69), and 3,4-diaminobenzophenone (DABP) (70), were found to be effective matrices later, 3-HPA remained to be the most popular matrix for nucleic acids. Glycerol was found to be a superior matrix for IR-MALDI at 2.94 μm , and was demonstrated to be able to

produce molecular ions for RNA transcripts as large as 2180 nucleotides (71). However, applications with this technique have been lacking so far.

Matrix additives have been proven to be important in MALDI of nucleic acids. Ammonium salts such as ammonium acetate (72), ammonium citrate, ammonium tartrate (60) and ammonium fluoride (73) were found to be useful additives which significantly suppress the adduction of metal cations. However, the molar concentration of ammonium salt in the matrix solution should be limited to about 10% of that of the matrix, as an excess amount of ammonium salt, when crystallized with matrix, reduces the absorption of laser energy and quenches the ionization of nucleic acids. Spermine as a cation exchanger, has also been used as an additive to several matrices (66,74), but not to 3-HPA (74). Other compounds, such as PA, various sugar (75) and PCA (68) have been added to 3-HPA to enhance its performance. However, the mechanism of such improvement remains elusive.

1.4.4 Sample Preparation

Small-scale sample preparation is usually performed manually, whereas an automatic nanoliter dispensing system is required for large scale high-throughput preparation. Several manual preparation methods are available for different matrices. The most common method is the traditional dried-droplet method, which is performed by spotting a droplet of the analyte-matrix mixture on the MS sample stage and drying the spot at room temperature. This method is modified for matrices that are soluble only in aqueous solvent, such as 3-HPA. The first step is to deposit a droplet of matrix in aqueous solution onto the sample plate and allow the matrix to crystallize at room temperature. The second step is then to add the

aqueous analyte solution onto the crystallized matrix. This partially re-dissolves the matrix crystals and re-crystallizes them together with analyte molecules. In general, the two-step dried-droplet method gives more reproducible results than the traditional one. Another method called thin-layer matrix preparation is reported, ATT matrix is dissolved in organic solvent to form a seed layer of thin homogeneous film, and then saturated ATT solution containing analyte is added to form the second layer of thin film. The thin layer preparation method can generate highly reproducible MS results (76).

For high-throughput MALDI-MS, sample spots have to be miniaturized to improve homogeneity of crystal distribution. The surface of the MALDI sample plate (metal or silicon chip) is usually coated with highly hydrophobic material, with precisely located and miniaturized sample pads (e.g. $200 \times 200 \mu\text{m}^2$) that are hydrophilic or less hydrophobic than their surroundings. Matrix solution can be transferred to the MALDI sample plate by high speed serial dispensing using a nano-dispensing device such as a piezoelectric pipette or solenoid dispensing valve. The aqueous matrix solution initially occupies an area larger than the sample pad. While the solvent evaporates, the droplet size shrinks, and eventually, crystallization of the matrix happens only inside the sample pad. The concentration of the matrix is optimized so that the crystallized matrix covers the entire sample pad. The aqueous analyte solution, usually from a microtiter plate, can be transferred to the MALDI sample plate preloaded with matrix by parallel dispensing using a pintool device with an array of precisely mounted pins (77). The analyte solution partially re-dissolves the 3-HPA matrix. The droplet shrinks again toward the hydrophilic pad while solvent evaporates, and then re-crystallizes to cover the pad. This miniaturized sample spot can provide more

homogeneous sample distribution thus yield more reproducible MALDI results.

Another important aspect of sample preparation is to desalt nucleic acid samples before mixing them with matrix, as sodium and potassium cations form multiple adducts with the negatively charged phosphate backbone easily and degrade the quality of mass spectra. High concentration of salt can also affect matrix crystallization and prevent successful ionization of analyte molecules. In biological assays, the use of sodium or potassium salt should be avoided as much as possible, or replaced with ammonium salt as it is volatile and usually does not form adducts to MALDI ions. As sodium and potassium are ubiquitous, desalting before MALDI-MS is always recommended. This can be achieved by using chromatographic pipette tips (78), solid-phase extraction (79) or by adding cation exchange resins to the sample solution (58).

1.4.5 Mass Spectrometry Instrumentation

Nucleic acids are less stable compared to proteins and peptides. The N-glycosidic bond linking the base and the pentose sugar is relatively weak, especially in DNA. Therefore, in MALDI, base loss, and subsequent backbone fragmentation could happen (80). Usually, MALDI ion source is coupled with a time-of-flight (TOF) mass spectrometer. The instability of DNA ions could cause significant amount of post-source decay (PSD) and result in reduced sensitivity and resolution if a reflectron TOF is used. Therefore, a linear TOF is commonly used to minimize this effect. However, for small oligonucleotides, as those generated by base-specific cleavage, PSD does not seem to be extensive and a reflectron can be used to improve mass resolution (76,81). It is also expected that high performance can be achieved in a quadrupole-TOF if collisional cooling is employed. The quadrupole must be

designed to transmit high m/z ions generated by MALDI. When collisional cooling is effectively used, MALDI-generated DNA ions can be stable for a significantly long period (milliseconds), which allows for high resolution detection in a fourier transform ion cyclotron resonance (FTICR)-MS (82), or in the newly developed Orbitrap with a MALDI ion source.

1.5 Primer Extension Assay

Primer extension assay is one of the major strategies employed to characterize the sequence changes using MALDI mass spectrometry. Primer extension reactions are designed for genotyping of known polymorphic sites and determining the levels of gene or allele expressions.

1.5.1 SNP Genotyping

To genotype known polymorphic site, the primer extension assay combined with MALDI-MS analysis is widely agreed as a powerful platform. The assay is usually performed in three major steps. The first step is PCR amplification of the region which contains the SNP of interest. Second, phosphatase is added to inactivate the remaining dNTPs. Third, a primer extension reaction is performed using a specific genotyping primer positioned with its 3'-end adjacent to the polymorphic site. This reaction generates allele-specific primer extension products with different masses which can be interrogated by MALDI-MS for genotype determination. There are two different versions of primer extension assay depending on the type of dNTP and ddNTP used in the primer extension reaction.

1) Primer extension with variable number of nucleotides (Fig. 2a)

In this primer extension reaction, a mixture of dNTP and ddNTP is used. The content of the mixture is selected so that the primer extension stops at the polymorphic site for one allele and at another site downstream for the other allele. Therefore the SNP resulted in allele-specific primer extension products with different lengths; one extended by a single base, while the other by two or more bases. The large mass difference between the products makes identification by MALDI-MS a trivial task, even for heterozygous samples. Braun et. al. first used this strategy to detect mutations in the CFTR gene (83). And it was later dubbed as homogenous MassEXTEND (hME) by Sequenom for multiplexed genotyping of different gene loci. The main drawback of this strategy is that in multiplexing, where multiple primers are used in a single reaction to target many different SNPs, careful planning is required to avoid peak overlapping. Even so, the level of multiplexing is moderate (up to 12-plex), and only those using the same type of dNTP and ddNTPs can be multiplexed.

2) Single-base primer extension (Fig. 2b)

This type of primer extension uses only ddNTPs. Therefore, the genotyping primer is only extended by a single base onto the polymorphic site. The genotype is differentiated by the mass of the ddNTP incorporated (84). This strategy is relatively simple as a common ddNTP mixture is used for all different types of SNPs. When the single-base extension is combined with different primers with various

5'-overhang (addition of non-templated bases), a high level of multiplexing can be achieved. Ross et. al. first demonstrated this potential in 1998 (85). However, as most primer extension reactions are measured in linear time-of-flight mass spectrometers to minimize the effect of post-source decay of MALDI ions, mass resolution is usually low and it is not always possible to resolve an A/T polymorphism where the mass difference between the two alleles differs by only 9 Da. This limitation can be overcome by using a modified ddATP or ddTTP. This modified ddATP or ddTTP must have mass value larger than ddGTP by at least 16 Da, and can be incorporated efficiently by DNA polymerase. Recently, Sequenom has used such strategy in their iPLEX genotyping design although detailed modification has not been published (86).

Figure 2. A comparison of the primer extension assays. (a) In the primer extension assay with variable number of nucleotides, at least one dNTP is used. The extension products differ by at least one base, a larger mass shift which can be easily resolved. (b) In single-base primer extension assay, only ddNTPs are used and the primers will always extend by one base no matter what the sequence is. The natural mass differences of the extended bases are used for SNP genotyping.

The MS-based primer extension approach was widely applied in a variety of fields that require the study of SNPs (87,88). Jaremko et al. reported successful SNP genotyping of DNA samples isolated from formalin-fixed paraffin-embedded tissues (FFPET) (89). FFPE materials are usually not suitable for molecular disease approaches, as they tend to yield degraded DNA, but they perfectly meet the need for long-term clinical research. With MALDI-TOF MS and Taqman assisted SNP genotyping approaches, the FFPET can be genotyped in a relatively high throughput format. Horn et al. further proved the reliability of the approach (90). Five DNA samples were extracted from FFPE tumor specimens using both phenol-chloroform extraction and commercial extraction kit. Thirty-one SNPs from 25 genes

were analyzed using the multiplex primer extension assay and 64 FFPE tumor specimens were compared with matched germline DNA samples and the genotyping result appeared to be reliable.

The MALDI-TOF MS-based approach is also used for oncogenic human papillomavirus (HPV) genotyping, providing an economical and high-throughput platform for the cervical screening programs and HPV vaccination programs. Soderlund-Strand et al. applied the primer extension assay to the detection of 14 oncogenic HPV genotypes (91). The MS-based genotyping result was compared with reverse dot blot hybridization (RDBH) result for 532 cervical cell samples. After comparison, it was found that the MS-based primer extension assay was sensitive enough to detect not only all samples that appeared positive under RDBH, but five more cases of cervical disease that RDBH had missed as well. All the 20 discrepancies were finally proved to be positive by the MS-based primer extension assay.

MALDI-TOF MS based primer extension assay also had good performance in the analysis of mitochondrial DNA coding region SNP (mtSNP) variation. Cerezo et al. compared MS-base primer extension assay with other techniques for the mtSNP genotyping and found that to genotype large amount of SNPs in large collections of samples, MS-based techniques appeared to be more suitable (92). Thompson et al. also employed the MS-based primer extension assay for SNP genotyping of 635 trees from 15 populations and the result was used as indirect evidence of realized F1 fertility for *populous* hybrids (93).

Thomas et al. employed the primer extension method for high-throughput genotyping to analyze oncogenic mutation in human cancer (94) and this method is also useful in other mutation profiling (95,96).

Huebner et al. optimized the MS-based SNP genotyping method to prevent the genotyping errors caused by tri-allelic SNPs (97). Sasayama's group reported an alternative MS-based genotyping approach using Lu(III) ion as molecular scissors(98). Compared with primer extension assay, their approach is independent on the sequence investigated thus is able to genotype not only SNPs but indels as well.

1.5.2 Genome-Wide Association Study

For large-scale population SNP analysis, DNA samples are commonly pooled as a mixture for the cost consideration, and the allele frequency of a certain SNP is usually calculated from the ratio of the alleles. The SNP frequency analysis can be used in the genome-wide association study, aiming to identify genes involved in certain diseases. Association is a statistical statement about the co-occurrence of alleles and/or phenotypes. The genome-wide association studies usually examine most, if not all, of the genes from different individuals of one particular population. Allele frequencies of all genes from the patients group (with certain disease) are compared with the control (without disease) group to identify the association between SNPs in different alleles and the disease. A large number of SNPs, most of which are in non-coding regions, are found to be associated with certain diseases in this way (99,100).

Kammerer et al. performed a large-scale of association study in order to identify genes that influence the risk to develop breast and prostate cancer (101). More than twenty-five thousand SNPs from sixteen thousand genes were screened and *ICAM* gene region was finally found to have the most significant association with the breast and prostate cancer risk.

Suttner et al. reported an association study between *TBX21* gene and asthma (102). In their work, the whole *TBX21* gene was resequenced and screened for polymorphisms. Forty three polymorphisms were identified using MALDI-TOF mass spectrometry and three of them were found to be associated strongly with asthma. These polymorphic alleles are combined with SNPs within the *HLXI* genes, which had been shown to have association with asthma in previous study. The combination resulted in the increased risk to develop asthma.

Abel and coworkers analyzed pooled DNA with genome-wide SNPs using the primer extension assay and proved its usefulness for discovering candidate susceptibility genes for common diseases (103). It is unlikely to get the unambiguous identification of the related gene or genetic variations for those common diseases using this genetic analysis only, but it is possible to locate the region of most potential for further experiments to identify the genes and variations responsible for the disease susceptibility. Osteoarthritis, for example, is a common disease with which neither traditional family-based linkage studies nor candidate gene studies can identify genes of significant association. Genome-wide association approach was performed and only a small amount of SNPs were selected for genotyping from thousands of SNPs. Finally, an intronic SNP in a gene, which is on chromosome 13 and encodes an unknown new protein with a calponin homology domain, was considered as the marker with the strongest association for risk of knee osteoarthritis (104).

Similar methodologies are also adopted for pathogenic organisms' research. Volkman et. al. genotyped about 47,000 SNPs across the genome of the malaria parasite *Plasmodium falciparum*(105). Their study generated a large data set which is of value for genome-wide diversity analysis. Herring et al. performed comparative genome sequencing of *Escherichia*

coli for bacterial evolution study and the MS based-primer extension assay was applied for mutation validation (106).

1.5.3 Gene Expression Analysis

The primer extension assay works well not only in SNP genotyping research, but also in quantitative gene expression studies. For quantitative gene expression analysis, primer extension assay is combined with competitive real-time PCR, followed by MALDI-MS analysis. Isolated RNA is used for reverse transcription and a synthetic DNA of about 90 nucleotides in length, referred as the “competitor”, is spiked into the cDNA product. This competitor differs from the cDNA by only one base in the center region and is used as an internal standard. The competitor internal standard serves as an artificial allele so that the method used for allele frequency analysis can be used here for gene expression analysis. The competitor and cDNA is co-amplified by PCR, after which shrimp alkaline phosphatase (SAP) is added to inactivate the unused nucleotides. Primer extension assay is then performed using the PCR products from the competitor and cDNA as templates. The primer extension products are submitted to MALDI-TOF analysis and the related mass signals are resolved to calculate the peak area ratios (107).

Rode et al. compared this MS-based primer extension assay with the microarray assay, which is a commonly used method for gene expression analysis, and tested the results from both MassARRAY and microarray against quantitative real-time PCR (qRT-PCR), which is widely agreed as a standard method for single gene expression quantification (108). Their study showed a high rate of concordance among the MS-base primer extension assay,

microarray and qRT-PCR in most of the data. The MS-based primer extension assay is proved suitable for prokaryotes gene expression research and became a better choice than microarray when analyzing a limited set of genes instead of global gene expression study. The MS-based primer extension assay is also advantageous over microarray in the low requirement of sample amount.

Duffield et al. combined the MS-based primer extension and the molecular affinity isolation for multiple mRNA levels determination (109). The mRNA levels of 4 genes with significant variation were determined simultaneously in their assay. Biotinylated ddNTPs were used in the single base extension which not only enlarged the mass difference between the extension products of the samples and the competitors, but also facilitated a purification step before MS analysis. Unextended primers, which had potential to overlap with other extended products, were all removed, releasing plenty of space in the spectrum to enable a higher level of multiplexing.

1.5.4 Non-Invasive Prenatal Diagnosis

The primer extension assay combined with MS technique is of great potential in risk-free, non-invasive prenatal diagnosis. Ding et al. introduced the use of single allele base extension reaction (SABER) and mass spectrometry for prenatal diagnosis in 2004 (110). Circulating fetal DNA can be detected from the maternal plasma. However, if the standard primer extension assay is performed, the maternal allele can always form an overwhelming background. The SABER protocol is able to extend the paternal-specific fetal allele, avoiding the interference caused by the background of maternal allele.

Daelemans et. al. analyzed candidate imprinted genes in human placenta and the allele-specific gene expression (111). Sequenom and Illumina assays, two typical allele-specific high-throughput technologies, are employed and both of them are of enough sensitivity for imprinting and allelic bias studies.

Lo et al. applied the MS-based primer extension assay to analyze the ratio between alleles of a SNP in *PLAC4* mRNA in order to perform non-invasive diagnosis of fetal trisomy 21 (112). The same technique can also be used in non-invasive prenatal diagnosis of trisomy 13 and 18. Further development of noninvasive prenatal diagnosis depending on MS-based primer extension technology for gene expression analysis has been reviewed elsewhere (113).

1.6 Base-Specific Cleavage Assay

Base-specific cleavage assay, combined with MALDI-MS, could be used for discovery of unknown polymorphisms and characterization of modifications.

1.6.1 SNP Discovery

In the past two decades, a number of different methods have been developed to achieve base-specific cleavage of both large DNA and RNA products using enzymatic approach as well as chemical means (114-118), in order to make it possible for MALDI-TOF MS measurement. Within all these methods, the most prominent approach introduces an intermediate RNA transcription step after PCR amplification of the target sequence region. Compared with direct cleavage of DNA product, this assay has several advantages. First of all, the cleavage of double-stranded DNA will certainly result in an overlapping cleavage patterns from each of the forward and reverse strand. After the transcription step, the single-stranded RNA product is ready for cleavage, avoiding the elaborate strands separation steps. Second,

the transcription step could further amplify the analyte, thus make it possible to add ion exchange treatment and dilution, which are important conditioning steps before MALDI-TOF analysis. Third, RNA is more stable than DNA under standard UV-MALDI condition as RNA is less prone to depurination.

The concept of base-specific cleavage has also been combined with MALDI-TOF MS (117,119) to discover new SNPs and detect mutations. In this process, there are two major enzymatic approaches to achieve base-specific cleavage of RNA transcripts. The first one uses base-specific RNase, for example, the use of RNase T1 could yield G-specific cleavage in normal working condition of the enzyme. This G-specific cleavage reaction could also be used in the reverse strand of the target region. The second method (Fig. 3) uses a mutant RNA polymerase – T7 or SP6 R&DNA polymerase, which can incorporate some of the dNTPs in transcription, combined with pyrimidine-specific RNase, for example RNase A. When mixed with normal transcript, RNase A would cleave after every rC and rU residue, resulting in more frequent fragmentation of RNA and loss of sequence information. However, as the mutant transcriptase could incorporate certain kinds of dNTPs (dCTP, dTTP, dUTP), base-specific cleavage can be achieved when either rC or rU is fully replaced by dC or dT/dU, respectively. In one reaction, dCTP is used instead of rCTP during transcription, which results in the U-specific cleavage of the transcript by RNase A, and in the other reaction, rUTP is fully replaced by dTTP or dUTP, and the transcript will finally be C-specifically cleaved. Considering that T7 promoter tag can be added to either the forward primer or the reverse primer in the PCR amplification step, one dsDNA sequence can actually be cleaved in both forward and reverse strands in four separate reactions. The cleavage products will be

conditioned for mass spectrometric analysis by addition of ion exchange resin and the supernatant will be transferred onto a silicon chip containing a 16x24 array of matrix spots, which can be automatically scanned in a MALDI mass spectrometer.

The combination of base-specific cleavage and MALDI-TOF MS offers extremely high sample throughput for large-scale comparative sequence analysis. It takes only a few seconds to measure each sample spot. If there is a sequence change in the target region, no matter what it is, up to 10 “observations” can be found in the mass spectra from the four base-specific cleavage reactions, two reactions will have mass shift, one reaction will introduce a cleavage site, and the last one will lose a cleavage site. If the sequence change does not alter a cleavage site, a mass peak shift will appear in one spectrum. The mass shift is actually two observations: the disappearance of a theoretical peak and the appearance of a new peak. If the sequence change alters a cleavage site, three changes will appear in one spectrum: if a new cleavage site is introduced by the sequence change, a peak with high mass will be replaced by two peaks with lower masses; if a cleavage site is abolished by the sequence change, two peaks with lower masses will disappear and a peak with higher mass will appear. For example, there is a sequence change of A to G, in the forward direction transcript, it is still an A to G change and in both rC and rU specific cleavage reaction, there is only a mass shift of 16 Da. While in the reverse direction transcript, it becomes a U to C change. In the rU specific cleavage reaction, one cleave site will disappear and in the rC specific cleavage reaction, a new cleavage site will be introduced. This comparative sequencing with multiple “observations” shows great advantage for diagnostic applications.

Similar to the identification of an unknown protein by the peptide mass fingerprinting

(PMF) approach (120-122), theoretical RNA sequences of a certain size are *in silico* digested with base specific enzyme like RNase A or RNase T1, and the molecular weight patterns for the predicted cleavage products are collected in the database. The mass spectra for the cleavage products from an unknown RNA sequence can be searched against the existing whole prokaryotic genomes or RNA FASTA sequence databases to find the best match or even identification using certain software such as “RNA mass mapping” (RMM) developed by Matthiesen and Kirpekar (123).

For SNP screening from a known sequence, a software package named “RNaseCut” was developed by Krebs et. al. (118). It can be used to design primers for those regions of high SNP discovery score and validation of individual SNPs. When a target sequence is given, the software first determines the best position for PCR amplification, in which more SNPs are theoretically able to be detected and validated. Then the transcribed RNA product is *in silico* digested to generate reference mass peak patterns. When the experimental observed spectra have differences compared to the reference, the software will calculate the possible solutions for a SNP which can cause the observed variations (124). A more sophisticated algorithm for SNP discovery has also been developed (116) but the software itself remains proprietary.

1.6.2 Identification of tRNA

It is difficult to get high resolution separation of transfer RNAs (tRNAs) using traditional methods because of the high level of similarity in their secondary and tertiary structure. Using base-specific cleavage & MALDI-MS assay, tRNAs are digested with RNase T1 and the generated fragments are found to contain unique masses for each individual tRNA. In this way, molecular mass peak patterns can be used as fingerprint for detection and identification of individual tRNAs from a total tRNA pool. This protocol does not require separation or purification step of the tRNAs sample thus is of relatively high efficiency (125).

1.6.3 RNA Modifications Analysis

Mass spectrometry nowadays serves as a powerful technology for the analysis of posttranscriptional modifications of different types of RNA (126-128), especially for those modifications that are not detectable by the RT-primer extension assay and the chromatographic assay (e.g. m^5U , m^5C , m^6A etc.) (129). Kirpekar *et al.* was the first to apply the base-specific cleavage and MALDI-MS assay to screen the 5S rRNA, in order to identify posttranscriptional modifications (130). The whole approach was of high speed and sensitivity while the procedure was relatively simple with only one purification step using ion-exchange resin. Suppression of the mass signal for some fragment might be observed but the total sequence coverage remained satisfactory. However, this approach lacks the ability to identify mass-neutral modifications. For example, the replacement of uridine by its structural isomer, pseudouridine will not lead to any mass difference and thus is not detectable by mass spectrometry. Several methods were reported in order to identify the pseudouridine, as its presence accounted for large part of the RNA modification in Eukarya.

Mengel-Jorgensen and Kirpekar generated a 53 Da mass increase to pseudouridine through cyanoethylation (131). Durairaj *et al.* utilized N-cyclohexyl-N'-β-(4-methylmorpholinium) ethylcarbodiimide (CMC) derivatization and alkaline treatment which introduced a 252 Da mass shift to the pseudouridine and also make it detectable (132).

Auxilien *et al.* used MALDI-MS to identify the target region of m⁵U methyltransferases by pinpointing the base-specific cleaved methylated RNA transcript. Target specificities of several m⁵U methyltransferase were analyzed and some shifts of target specificity were observed in archaeal RNA m⁵U methyltransferase (133). Using MALDI mass spectrometry, Goll *et al.* found that the target of human DNA methyltransferase-2 (DNMT2), which should be cytosine in the anticodon loop, was actually an RNA instead of DNA (134). This RNA was identified to be aspartic acid transfer RNA (tRNA^{ASP}). On the basis of this observation, they suggested that eukaryotic DNA cytosine methyltransferase might be derived from an RNA methyltransferase instead of a prokaryotic DNA methyltransferase.

1.6.4 DNA Methylation Analysis

Base-specific nucleic acid cleavage combined with MALDI mass spectrometry technique has also been used for DNA methylation analysis. DNA methylation occurs by adding a methyl group to the number 5 carbon of the cytosine pyrimidine ring or the number 6 nitrogen of the adenine purine ring. DNA methylation in somatic tissues is typically observed at the cytosine base in a CpG dinucleotide context (135,136), while non-CpG methylation occurs mostly in stem cells (136-138).

The target DNA is first treated with bisulfite, which can mediate a deamination of

unmethylated cytosines to uracils, while the methylated cytosines stay unchanged. After the bisulfite treatment, the target DNA is PCR amplified before used for *in vitro* transcription. The generated single-stranded RNA product will then undergo RNase A mediated base specific cleavage procedure by replacing either rCTP or rUTP with their dNTP counterparts. The CpG methylation could be detected in both the forward and the reverse strand. For forward strand as an example, in the C-specific cleavage reaction, methylated cytosines remain unchanged and will be cleaved by the RNase while the unmethylated cytosines will finally be transcribed using dUTPs and these cleavage sites are abolished. For the reverse strand, the methylation site will appear as G/A variation and the fragments containing these sites will have a mass shift of 16 Da. By analyzing these mass shift, methylation could be easily detected and by comparing with the signal intensity of mass peaks from fully methylated and unmethylated templates, quantitative analysis of methylation is possible (135,139-142).

1.6.5 MLST for Bacterial Studies

At the end of the last century, in order to discriminate the strains of infectious agents which caused disease, there were two different approaches. The first approach identified individual loci or the uncharacterized genomic regions which were variable among bacterial populations. Several methods were developed based on this approach for the determination of the relatedness of pathogens. These included pulsed-field gel electrophoresis (PFGE), ribotyping and PCR-based fingerprinting. PFGE was widely considered as “gold standard” even though this technique was unable to type many strains as a result of the DNA

degradation during the electrophoresis process. The second approach was based on the characterization of variations in the sequence of many loci, which were slowly accumulating in the microbial population. Multilocus enzyme electrophoresis (MLEE) was developed based on this approach. Because it can analyze several loci at the same time,. MLEE was able to achieve high level of discrimination. However, like many other typing methods, the result of the MLEE was not portable between different laboratories thus had difficulty in comparison (143).

Multilocus sequence typing (MLST) was a comparative sequence typing approach on the principle of MLEE, for the typing of multiple loci. It was first introduced by Martin C. Maiden and the coworkers in 1998. In their research, MLST was used to characterize the population structure of bacterial isolates with the analysis of DNA sequences of internal fragments of multiple housekeeping genes (143-145).

The proposal of MLST has improved knowledge of bacterial evolution and population biology (146-149) while at the same time made the high-throughput determination of nucleotide sequence much more available and cost-efficient (150,151), and the most significant advantage of MLST over other molecular typing approaches mentioned above was that it could be applied to almost all bacterial species even those that are very difficult to cultivate, and the sequence data were highly unambiguous and portable between laboratories— all the materials for MLST could be exchanged between labs and the primer sequences and the protocol could be shared via internet thus for the first time make it possible to establish a global database for any individual species on the World-Wide Web site. The MLST database covers 27 species today, and will continuously expand in the future.

As the typical sequence-based typing methodology (152), MLST could be combined with base-specific cleavage assay and MALDI-TOF mass spectrometry for microbial typing (153). This nucleic acid-based typing approach did not need any prior targets information, which was critical for traditional microbial typing methods that were derived from *in situ* hybridization (154). The 16S small ribosomal RNA gene, which was commonly used for bacterial identification (155), was PCR amplified for base-specific-MS analysis. The data acquired were searched against the database containing the *in silico* digest result of existing 16S rRNA sequences. The whole process took only a few hours and was compatible with both bacterial isolates and uncultured clones (156-160).

1.6.6 Pathogenic Viruses Study

The base-specific cleavage and MALDI-MS approach has already been used for hepatitis B virus (HBV) genotyping (161). Ganova-Raeva's group compared the mass spectrometry-based approach with the traditional sequencing for HBV genotyping and proved that the new approach was much more cost-effective but of equal reliability. For large scale analysis which require high throughput, the cost for base-specific cleavage & MALDI-MS assay can be much lower than that for sequencing.

This new approach also has potential in the rapid identification and mutation surveillance of other pathogenic viruses like influenza. Commonly used methods for rapid diagnosis of influenza are usually based on immunoassays using antigen-specific antibodies (162), or real-time reverse transcription PCR (RT-PCR) of the M gene to identify viral species (163), followed by a second round of RT-PCR assay for HA and NA genes to determine the

subtypes (164,165). Compared with these methods, the approach based on base specific cleavage and MALDI mass spectrometry not only has advantage in the throughput and speed, but also in the ability for the identification of new emerging genetic variants with mutations in the drug-resistant mutation site (166,167).

1.7 Influenza A Virus

Influenza is a segmented negative-strand RNA virus, which belongs to the family of *Orthomyxoviridae*. The common symptoms of influenza infection are fever, chills, cough, myalgias, headache, and sore throat. In more serious cases, influenza causes pneumonia, which can be fatal, particularly among the young and the elderly (162,168,169).

Influenza pandemics killed millions of people worldwide in human history. Spanish flu pandemic in 1918 killed more than 25 million people within one year. The 1957 Asian flu resulted in 2 million deaths. More than 1 million people died worldwide from the 1968 Hong Kong flu. In March and early April 2009, at the end of flu season in the northern hemisphere, a new swine-origin influenza virus emerged in Mexico and the United States, and spread globally by human-to-human transmission (170-172).

Influenza viruses are classified into three immunological types, A, B, and C, depending on the antigenic properties of their different nucleoproteins and matrix proteins. The influenza A virus has caused all of the major pandemics in humans and has therefore been more extensively investigated than the B and the C types. Influenza A virus (Fig. 4) infects a wide variety of mammalian and avian species in addition to humans. It is believed that influenza A viruses are transported around the world by migratory birds and air travel.

The species specificity of a particular viral strain presumably arises from the binding specificity of its HA for cell-surface glycolipids.

Hemagglutinin (HA) is an elongated trimeric transmembrane glycoprotein, playing a central role in both the viral infection process and in the immunological measures and countermeasures taken in the continuing biological contest between host and parasite. HA is largely responsible for stimulating the production of the antibodies that neutralize the virus. Consequently, the different influenza virus subtypes arise mainly through the variation of HA. Two distinct mechanisms of antigenic variation have been observed in influenza A viruses: antigenic shift and antigenic drift.

In antigenic shift, the gene encoding one HA species is replaced by an entirely new one. It is thought that these new viral strains arise from the reassortment of genes among animal and human influenza viruses. Antigenic shift is responsible for influenza pandemics because the human population's immunity against previously existing viral strains is ineffective against the newly generated strain. Evidently, these viruses had retained the genetic traits responsible for their virulence in humans.

Antigenic drift occurs through a succession of point mutations in the HA genes, resulting in an accumulation of amino acid residue changes that attenuate the host's immunity. This process occurs in response to the selective pressure brought about by the buildup in the human population of immunity to the extant viral strains. HA varies in this manner by an average of 3.5 accepted amino acid changes per year.

Neuraminidase (NA) is another major surface protein of influenza virus. NA is a tetrameric transmembrane glycoprotein which catalyzes the hydrolysis of the linkage joining

a terminal sialic acid residue to a D-galactose or a D-galactosamine residue. NA probably facilitates the transport of the virus to and from the infection site by permitting its passage through mucin and preventing viral self-aggregation. Antigenic variation in NA also occurs but this has less immunological consequences.

Adapted from Subbarao and Joseph, 2007

Figure 4. Shown above is the schematic structure of Influenza virus(173). Influenza belongs to the family of *Orthomyxoviridae*. As shown in the figure, the envelope of influenza virus consists of two glycoproteins, hemagglutinin and neuraminidase. According to the similarities of these two glycoproteins, Influenza A viruses could be classified into subtypes (H1N1, H3N2, H5N1, etc.)

Influenza A viruses are classified into subtypes according to the similarities of their HA and NA. There are 16 known subtypes of HA (H1~H16) and 9 of NA (N1~N9) that occur in mammals and birds. Avian virus subtypes occur in nearly all combinations, while only a

few combinations have been found in humans.

Humans are rarely infected by avian flu viruses and such viruses do not appear to be transmitted between humans. However, phylogenetic studies indicate that pigs (swine) and birds can exchange influenza viruses as can pigs and humans. This suggests that pigs serve as “mixing vessels” for the creation of new pandemic flu viruses(174).

Commonly used methods for rapid influenza viruses diagnosis rely on antigen-specific antibody assays(162), or real-time RT-PCR analysis of the matrix (M) gene for identification of the viral species(163) followed by HA and NA subtype-specific RT-PCR assays for determination of the viral subtypes (164,165). However, these methods lack the ability to identify reassortments and new emerging genetic variants as there are many HA and NA subtypes with significant intra- and inter- subtype sequence variations. Hofstadler(175) described a strategy, termed TIGER (Triangulation Identification for the Genetic Evaluation of Risks), for detection and characterization of microorganisms based on MS-derived base composition signatures obtained from PCR amplification of broadly conserved regions of the microbial genomes. First, the masses of a PCR product (positive and negative strands) are measured using electrospray ionization (ESI). The base compositions of the positive and negative strands can be deduced based on the constraint that they are complementary. A mutation can be detected by a mass shift, therefore, a composition change. This strategy, although conceptually simple, requires high mass resolution (>15000) and high mass accuracy (<20ppm) to deduce mono-isotope masses of PCR products. This limits the measurable length of PCR products to less than 150 base-pair. This method also lacks the ability to locate the position of a point mutation within the analyzed sequence. The

requirement of using ESI also demands for rigorous purification and desalting of PCR products which is usually difficult.

Amantadine and rimantadine (Fig. 5) used to be effective drugs for influenza as M2 ion channel blockers. However, both the seasonal H3N2 flu and the H1N1 Swine flu in 2009 were found to have acquired Ser31Asn mutation in the M2 protein, which resulted in their resistance to these adamantane-derived drugs(176,177). Neuraminidase inhibitors oseltamivir (tamiflu) and zanamivir (Relenza) are widely used today instead of amantadine and rimantadine for the treatment of influenza A infections.

Figure 5. Structure of amantadine (a) and rimantadine (b). Amantadine and rimantadine were used as antiviral drug by blocking the M2 ion channel. In the 2008-2009 flu season, seasonal H3N2 and pandemic H1N1 swine flu were found resistant to these drugs. After that, they were no longer recommended for influenza A infection treatment.

Alarmingly, a mutation His274Tyr has been found in neuraminidase in seasonal influenza A/H1N1(178), which has rendered oseltamivir (Fig. 6) useless, and the percentage

of resistant cases are still increasing. Some other mutations, such as Arg292Lys, Glu119Val and Asn294Ser, were found in seasonal circulating influenza A/H3N2 viruses. These mutations can also result in the resistance of oseltamivir.

Figure 6 Structure of oseltamivir (Tamiflu). Oseltamivir works as neuraminidase inhibitor and has been used to treat and prevent influenza A virus.

1.8 Research Objectives

1.8.1 Comparative Sequencing of H1N1 Influenza Virus

Influenza A H1N1 virus was the most common cause of human influenza during recent years especially in the 2009 flu pandemic. The isolated H1N1 swine influenza virus (SIV) was found to be a reassortment virus. It contains genetic elements from 4 different influenza viruses: two kinds of swine influenza viruses, one avian influenza virus and human influenza virus.

It is not easy to identify the reassortants or newly emerging genetic variants of influenza viruses by using common methods for influenza diagnosis, which mainly rely on probe techniques and RT-PCR. Here in this research, we adopted a comparative sequencing method

based on high throughput MALDI-MS for characterization and monitoring of influenza A H1N1 viruses as an alternative to direct genome sequencing.

We selected two influenza A H1N1 viral strains: WSN and PR8. Base-specific cleavage and MALDI-MS assay was used for comparative sequencing of the HA and NA genes. Other genes could be analyzed in the same way.

1.8.2 Surveillance of Genetic Drift

Genetic drift of influenza virus HA and NA genes might result in the resistance of the viruses to drugs. These genetic drifts need to be monitored in an efficient and effective way. Compared with direct genome sequencing, base-specific cleavage & MALDI-MS assay is more time and cost efficient, and has an advantage in the throughput.

Sequence changes in HA genes and NA genes may result in the genetic drift of the influenza virus. The comparative sequencing method we applied in this research, is based on MALDI-MS combined with base-specific assay, which depends on multiple verifications for each sequence change. This approach is used for monitoring of mutation sites in the viral genome.

Chapter Two. Materials and Methods

2.1 Materials and Reagents

All the common reagents that were used in our experiments were purchased from Bio-Rad, Sigma Aldrich, Invitrogen, Fluka, unless otherwise stated. The enzymes were all bought from Qiagen, Fermentas and Epicentre.

2.1.1 Virus

A/WSN/33 (H1N1)	ATCC
A/PR/8/34 (H1N1)	DSO

2.1.2 RT-PCR

RNAEasy Mini Kit	QIAGEN
------------------	--------

2.1.3 PCR

Primers	Sigma Aldrich
HiFi Platinum Taq Kit	Invitrogen
dNTP	Fermentas

2.1.4 DNA Analysis

1X TAE	20 ml 50X TAE dissolved in 980 ml of ddH ₂ O
--------	---

Ethidium Bromide Solution	1 mg of Ethidium Bromide desolved in 1L of ddH ₂ O
---------------------------	---

1.5% Agarose gel	1.5g Agarose (1 st BASE) in 100 ml 1X TAE
------------------	--

2.1.5 SAP Treatment

Shrimp Alkaline Phosphatase (SAP) Kit	Fermentas
---------------------------------------	-----------

2.1.6 *in vitro* Transcription

T7 R&DNA polymerase Kit	EPICENTRE
-------------------------	-----------

RNase free (DEPC) H ₂ O	Invitrogen
------------------------------------	------------

100mM dCTP	Fermentas
------------	-----------

100mM dUTP	Fermentas
------------	-----------

100mM rCTP	Fermentas
------------	-----------

100mM rUTP	Fermentas
------------	-----------

100mM rATP	Fermentas
------------	-----------

100mM rGTP	Fermentas
------------	-----------

dC mix	rU:rG:rA:dC =1:1:1:2.5
--------	------------------------

dU mix	rC:rG:rA:dU =1:1:1:2.5
--------	------------------------

2.1.7 RNaseA Digestion

10X 1mg/ml RNaseA

1mg RNaseA in 1ml 10mM NaAc, 100°C 15 min, add 0.1ml 1M Tris-Hcl

2.1.8 Resin

DOWEX 50W X 8-400
resion

0.5 g resin in 0.1 ml of 10mM ion-exchange
NH₄AC

2.1.9 MALDI

MALDI-MS

MALDI-TOF 4800 from ABI

50 mg/ml 3-HPA

5mg 3-HPA (Fluka) in 0.1ml 50% ACN
with 0.1 % TFA + 1.5 µl 2.5M Ammonium
citrate

50 mg/ml ATT

5mg ATT (Fluka) in 0.1ml 50% ACN with
0.1 % TFA

µFocus MALDI Plate 700µm (384 circles)

HST

2.2 Methods

2.2.1 Viral cDNA Preparation

2.2.1.1 Viral RNA Extraction.

Infected chicken allantoic fluid was used as starting material. RNeasy Mini Kit (QIAGEN) was used. Following the manufacturer's protocol, the viral RNA was finally eluted in RNase free water.

2.2.1.2 cDNA Synthesis

Eight microliters of extracted vRNA was mixed with 1 μ L of 10 mM dNTP mix and 1 μ L of unit12 primers before incubated at 65 °C for 5 minutes and then 4 °C for at least 1 minute. Two microliters of 10 $\times$ RT buffer, 4 μ L of 25 mM MgCl₂, 2 μ L of 0.1 M DTT and 1 μ L of RNase OUT (Invitrogen) was added to the RNA mixture and incubated at 42 °C for 2 minutes. One microliter of reverse transcription enzyme SuperscriptTM II (invitrogen) was added and the mixture was incubated at 42 °C for another 50 minutes. The RT enzyme was inactivated by treatment at 70°C for 15 minutes. Following the cDNA synthesis, 1 μ L of RNase H was used to remove the viral RNA from the product.

2.2.2 Base-Specific Cleavage Assay

2.2.2.1 Polymerase Chain Reaction (PCR) Amplification

A PCR reaction was performed to amplify the cDNA. A T7 promoter sequence was added at 5'-end of either the forward strand primer or the reverse strand primer.

The PCR mix per reaction was 7.2 μ l of ddH₂O, 1.0 μ l of 10X HiFi PCR buffer, 0.4 μ l of 50mM MgSO₄, 0.2 μ l of cDNA template, 0.4 μ l of 10 μ M forward primer, 0.4 μ l of 10 μ M reverse primer, 0.3 μ l of 10mM dNTP mixture, and 0.1 μ l of HiFi Platinum Taq enzyme (5U/ μ l, Invitrogen).

Each PCR reaction was subjected to one cycle at 95°C for 2 min; 35 cycles of a denaturing step at 95°C for 30 sec, an annealing step at 55°C for 30 sec, an extension step at 68°C for 1 min; and one cycle of an additional extension at 68°C for 7 min.

2.2.2.2 PCR product Checked by Agarose Gel Electrophoresis

The PCR product was checked by 1.5% agarose gel in 1X TAE buffer. The agarose was heated in microwave until boiled and cooled in room temperature for 10 min. The agarose was poured slowly into the gel tank (BioRad) and a comb as inserted. The agarose in the tank was left to set for one hour. 1X TAE buffer was poured into the gel tank to submerge the gel to 2 mm depth. The PCR product and 100 bp DNA ladder were loaded with 6X loading buffer and run under PowerPac (BioRad) at 90 V for 30 min. The agarose was then soaked in ethidium bromide solution for 30min before viewed with UV Gel Imager (BioRad).

2.2.2.3 Shrimp Alkaline Phosphatase (SAP) Treatment

The PCR products were treated with shrimp alkaline phosphatase (SAP) to deactivate remaining dNTPs in the solution: 1µl of SAP and 1µl of SAP working buffer was added in to 8µl of PCR products. The solution was treated at 37°C for 30 min and 85°C for 10 min.

2.2.2.4 *In vitro* Transcription

T7 R&DNA polymerase, which can incorporate both ribonucleic acid triphosphate (rNTP) and some deoxyribonucleic acid triphosphates (dCTP, dUTP and dTTP), was purchased from EPICENTRE. Each SAP treated PCR product was used in two separated transcription reactions.

Two dNTP/rNTP mixture were prepared instead of normal rNTP mixture and used for two *in vitro* transcriptions. The first mixture was “dC mix”, which is rATP, rTTP, rGTP, and dCTP; the second one was “dU mix”, which is rATP, rCTP, rGTP and dUTP.

The *in vitro* transcription mix per reaction was: 2µl of SAP treated PCR products, 1µl of

RNase free water, 1µl of 5X transcription buffer, 0.4µl of 100mM DTT, 0.4µl of “dC mix” or “dU mix”, and 0.2µl of T7 R&DNA polymerase. The 5µl transcription reaction was processed in 37°C incubator for four hours.

2.2.2.5 RNase A Digestion

After *in vitro* transcription, 0.55µl 10X RNase A (1mg/ml) was added into the transcription products. The cleavage reaction was performed at 37°C for 1 hour.

2.2.2.6 Purification of Cleavage Products

Before MALDI-TOF MS analysis, the cleavage products were desalted by Dowex® 50WX8-400 cation-exchange resin. Six milligrams of resin, which was resuspended in 10 mM ammonium acetate solution, was added into the cleavage products and was set in room temperature for 10 minutes.

2.2.3 MALDI-TOF MS Analysis

Approximately 0.3µl of the desalted products were transferred onto a 384-well plate and co-crystallized with 0.3µl of 50 mg/mL 3-HPA (or ATT) matrix. Spectra were acquired under reflector positive ion mode. Two thousand shots were acquired for each spectrum. Signal to noise threshold was set at 10, default calibration was used and laser intensity was fixed at 6500 arbitrary units.

2.2.4 *in silico* Digestion

The sequences of the transcription products were digested *in silico* using Mongo Oligo

Mass Calculator v2.06 (Fig. 7):

<http://library.med.utah.edu/masspec/mongo.htm>

For transcription with dC mix, the dC residues were represented as “N” with base anion mass of 110.03 and sugar add mass of -16. For transcription with dU mix, the dU residues were represented as “N” with base anion mass of 111.02 and sugar add mass of -16. Fragments with less than four residues were ignored and was not included in the theoretical mass list due to their poor specificity.

Mongo Oligo Mass Calculator v2.06

Sequence (5' --> 3')
 NAGUAAUANGANUNAUAGGGAGAUUNGAUAUAUUNNNAAGGAAAGUUNAUUGGNNNAANN/

Calculations

- ☐ molecular mass
- ☐ electrospray series
- ☐ CID fragments
- ☐ internal fragments
- ☐ base losses
- ☐ RNase T1 digest
- ☐ RNase U2 digest
- ☒ RNase A digest
- ☐ 5' exonuclease digest
- ☐ 3' exonuclease digest
- ☐ modified bases list

Options

- ☐ average mass
- ☒ monoisotopic mass
- ☐ negative mode
- ☒ positive mode
- ☐ DNA
- ☒ RNA
- ☐ allow undercut

Terminals

- ☒ 5' -OH
- ☐ 5' phosphate
- ☒ 3' -OH
- ☐ 3' phosphate
- ☐ 3' cyclic phosphate

Memories

select a memory location ----->

User residue 'N'

110.03 base anion mass
 -16 sugar add mass

Output

- ☒ close button
- ☒ output summary
- sequence by ☒ 3 ☐ 10 ☐ 50
- ☒ extended output
- ☒ show messages
- ☐ sorted output

written by Jef Rozanski (1999)
 visit the Nucleic Acids Masspec Toolbox

Figure 7. Mongo Oligo Mass Calculator v2.06. “base anion mass” would be set to 110.03 when dC residues were represented as “N”, 111.02 when dU residues were represented with “N” and 125.03 for dT reaction. Other parameters were set as shown above.

Chapter Three. Analysis of Two Genes from Two Viral Strains

3.1 Analysis of HA Genes from WSN & PR8

Four pairs of primers (Table 1) were designed to amplify 4 fragments in order to cover the full length of HA gene (Fig. 8). The sequences of HA gene for WSN and PR8 were from NCBI website:

(<http://www.ncbi.nlm.nih.gov/genomes/FLU/Database/multiple.cgi>)

F1: 14-507

F3: 845-1394

F2: 490-920

F4: 1221-1741

Figure 8. Shown above is the design of primers for HA gene. The HA gene is about 1750 bases in length, and the primers were all designed in the conserved region between WSN and PR8 viral strains. Four PCR reactions were performed in order to cover the full length of the HA gene. Each PCR reaction amplified a fragment with about 500 bases in length. The adjacent fragments would have an overlapped region.

Table 1. Primers for HA genes. HA gene was divided into four overlapping segments. For each of the segment, two different PCR reactions were performed. One reaction introduced a T7 promoter tag (-CAGTAATACGACTCACTATAGGGAGACAAAATGAAGGCAA) in the sense strand of the amplicon, and the other reaction in the antisense strand. E.g. HA+1F (with T7 promoter) and HA+1R were used to amplify the first segment, and the T7 promoter was added to the sense strand (+), while HA-1F (with T7 promoter) and HA-1R were used to amplify the same portion, but the T7 promoter was added to the antisense strand (-). As a result, both strands of this segment could be used for transcription in the later step.

Primer name	Primer sequence
HA+1F	CAGTAATACGACTCACTATAGGGAGACAAAATGAAGGCAA
HA+1R	GTCAGGCATAGCAAAATTTCTGTAAAAACTGC
HA-1F	CAGTAATACGACTCACTATAGGGAGAGCCATAGCAAAATTTCTG
HA-1R	CTACTGGTCCTGTTATGTGCACTTGC
HA+2F	CAGTAATACGACTCACTATAGGGAGACAGAAAATTTGCTATGGCTGAC
HA+2R	GACACTTCGTGTACACTCATGCAATTG
HA-2F	CAGTAATACGACTCACTATAGGGAGAGACACTTCGTGTACACTC
HA-2R	GTTTTTACAGAAAATTTGCTATGGCTGAC
HA+3F	CAGTAATACGACTCACTATAGGGAGATCCGGCATCATCACC
HA+3R	CCATTTCCGATTTCTTTGGCAATTATCTTT
HA-3F	CAGTAATACGACTCACTATAGGGAGACGATTTCTTTGGCAATTATTC
HA-3R	GTCCGGCATCATCACCTCAAACGC
HA+4F	CAGTAATACGACTCACTATAGGGAGATTTCTTTGGACATTTTGG
HA+4R	CCATTTAGAAACACATCCAGAAACTGATTTGCCCC
HA-4F	CAGTAATACGACTCACTATAGGGAGACCATTAGAACACATCCAG
HA-4R	GTTGATGATGGATTTCTTTGGACATTTGGAC

3.1.1 Fragment One to WSN HA

The sequence of PCR product (Fig. 9) from fragment 1 is:

```
5' -CCAAAATGAAGGCAAGACTACTGGTCCTGTTATGTGCACTTGCAGCTACAGATG
CAGACACAATATGTATAGGCTACCATGCGAACAACTCAACCGACACTGTTGACACAC
TACTCGAGAAGAATGTGACAGTGACACATTCTGTTAACCTGCTCGAAGACAGCCACA
ACGGGAAACTATGTAAATTAAAAGGAATAGCCCCACTACAATTGGGGAAATGTAACA
TCACCGGATGGCTCTTGGGAAACCCAGAATGCGACTCACTGCTTCCAGCGAGATCAT
GGTCCTACATTGTAGAAACACCAAACCTCTGAGAATGGAGCATGTTATCCAGGAGATT
TCATCGACTATGAGGAACTGAGGGAGCAATTGAGCTCAGTATCATCATTAGAAAGAT
TCGAAATATTTCCCAAGGAAAGTTCATGGCCCAACCACAACACACTCAACGGAGTAA
CAGTATCATGCTCCCATAGGGGAAAAAGCAGTTTTTTACAGAAATTTGCTATGGCTGA
C-3'
```

Figure 9. The figure shown above is the PCR products (~500 bps) and 100 base-pair DNA marker run on 1.5% agarose gel. The products appear near the fifth band (500bps) of the marker.

The PCR product was used for two *in vitro* transcriptions. In positive dC transcription reaction, the product is: (N=dC)

5' -NNAAAAUGAAGGNAAGANUANUGGUNNUGUUAUGUGNANUUGNAGNUANAGAUG
NAGANANAAUAUGUAUAGGNUANNAUGNGAANAANUNAANNGANANUGUUGANANAN
UANUNGAGAAGAAUGUGANAGUGANANAUUNUGUUAANNUGNUNGAAGANAGNNANA
ANGGGAAANUAUGUAAAUAAGGAAUAGNNNNANUANAUAUUGGGGAAAUGUAANA
UNANNGGAUGGNUNUUGGGAAANNAGAAUGNGANUNANUGNUUNNAGNGAGAUNA
GGUNNUANAUAUGUAGAAANANNAANUNUGAGAAUGGAGNAUGUUAUNNAGGAGAUU
UNAUNGANUAUGAGGAANUGAGGGAGNAAUUGAGNUNAGUAUNAUNAUAUAGAAAGAU
UNGAAUAUUUNNNAAGGAAAGUUNAUGGNNAANNANANANUNAANGGAGUAA
NAGUAUNAUGNUNNNAUAGGGGAAAAAGNAGUUUUUANAGAAAUUGNUAUGGNUGA
N-3'

The sequence could be *in silico* digested and the result is as follows:

pos	mass	3'>p	sequence
N1:U7	2218.315	2200.305	NNAAAAUp
G8:U19	3927.556	3909.546	GAAGGNAAGANUp
A20:U22	942.122	924.112	ANUp
G23:U25	1014.129	996.119	GGUp
N26:U28	902.106	884.096	NNUp
G29:U30	669.082	651.072	GUp
U31:U31	324.035	306.025	Up
A32:U33	653.087	635.077	AUp
G34:U35	669.082	651.072	GUp
G36:U40	1576.205	1558.195	GNANUp
U41:U41	324.035	306.025	Up
G42:U47	1921.252	1903.242	GNAGNUp

A48:U53	1945.274	1927.264	ANAGAU_p
G54:U64	3526.497	3508.487	GNAGANANAAU_p
A65:U66	653.087	635.077	AU_p
G67:U68	669.082	651.072	GU_p
A69:U70	653.087	635.077	AU_p
A71:U75	1632.217	1614.207	AGGNU_p
A76:U80	1560.210	1542.200	ANNAU_p
G81:U90	3197.445	3179.435	GNGAANAANU_p
N91:U101	3430.469	3412.459	NAANNGANANU_p
G102:U103	669.082	651.072	GU_p
U104:U104	324.035	306.025	U_p
G105:U112	2523.345	2505.335	GANANANU_p
A113:U115	942.122	924.112	ANU_p
N116:U125	3293.473	3275.463	NGAGAAGAAU_p
G126:U127	669.082	651.072	GU_p
G128:U133	1961.269	1943.259	GANAGU_p
G134:U140	2234.310	2216.300	GANANAU_p
U141:U141	324.035	306.025	U_p
N142:U143	613.070	595.060	NU_p
G144:U145	669.082	651.072	GU_p
U146:U146	324.035	306.025	U_p
A147:U151	1560.210	1542.200	AANNU_p
G152:U154	958.117	940.107	GNU_p
N155:U178	7708.090	7690.080	NGAAGANAGNNANAANGGGAAANU_p
A179:U180	653.087	635.077	AU_p
G181:U182	669.082	651.072	GU_p
A183:U186	1311.191	1293.181	AAAU_p
U187:U187	324.035	306.025	U_p
A188:U196	2988.443	2970.433	AAAAGGAAU_p
A197:U205	2772.364	2754.354	AGNNNNANU_p
A206:U210	1600.227	1582.217	ANAAU_p
U211:U211	324.035	306.025	U_p
G212:U219	2691.380	2673.370	GGGGAAAU_p
G220:U221	669.082	651.072	GU_p
A222:U226	1600.227	1582.217	AANAU_p
N227:U234	2539.340	2521.330	NANNGGAU_p
G235:U238	1303.165	1285.155	GGNU_p
N239:U240	613.070	595.060	NU_p
U241:U241	324.035	306.025	U_p
G242:U255	4545.644	4527.634	GGGAAANNNAGAAU_p
G256:U261	1921.252	1903.242	GNGANU_p
N262:U265	1231.158	1213.148	NANU_p
G266:U268	958.117	940.107	GNU_p
U269:U269	324.035	306.025	U_p

N270:U279	3213.440	3195.430	NNAGNGAGAU_p
N280:U282	942.122	924.112	NAU_p
G283:U285	1014.129	996.119	GGU_p
N286:U288	902.106	884.096	NNU_p
A289:U292	1271.175	1253.165	ANAU_p
U293:U293	324.035	306.025	U_p
G294:U295	669.082	651.072	GU_p
A296:U309	4457.642	4439.632	AGAAANANNAANAU_p
N310:U311	613.070	595.060	NU_p
G312:U317	2001.286	1983.276	GAGAAU_p
G318:U324	2306.316	2288.306	GGAGNAU_p
G325:U326	669.082	651.072	GU_p
U327:U327	324.035	306.025	U_p
A328:U329	653.087	635.077	AU_p
N330:U338	2924.404	2906.394	NNAGGAGAU_p
U339:U339	324.035	306.025	U_p
U340:U340	324.035	306.025	U_p
N341:U343	942.122	924.112	NAU_p
N344:U348	1576.205	1558.195	NGANU_p
A349:U350	653.087	635.077	AU_p
G351:U358	2635.369	2617.359	GAGGAANU_p
G359:U369	3654.515	3636.505	GAGGGAGNAAU_p
U370:U370	324.035	306.025	U_p
G371:U375	1632.217	1614.207	GAGNU_p
N376:U379	1287.170	1269.160	NAGU_p
A380:U381	653.087	635.077	AU_p
N382:U384	942.122	924.112	NAU_p
N385:U387	942.122	924.112	NAU_p
U388:U388	324.035	306.025	U_p
A389:U396	2659.390	2641.380	AGAAAGAU_p
U397:U397	324.035	306.025	U_p
N398:U403	1945.274	1927.264	NGAAAU_p
A404:U405	653.087	635.077	AU_p
U406:U406	324.035	306.025	U_p
U407:U407	324.035	306.025	U_p
N408:U419	3871.544	3853.534	NNNAAGGAAAGU_p
U420:U420	324.035	306.025	U_p
N421:U423	942.122	924.112	NAU_p
G424:U442	5918.815	5900.805	GGNNNAANNANANANAU_p
N443:U451	2924.404	2906.394	NAANGGAGU_p
A452:U457	1945.274	1927.264	AANAGU_p
A458:U459	653.087	635.077	AU_p
N460:U462	942.122	924.112	NAU_p
G463:U465	958.117	940.107	GNU_p

N466:U470	1520.194	1502.184	NNNAUp
A471:U485	4986.719	4968.709	AGGGGAAAAAGNAGUp
U486:U486	324.035	306.025	Up
U487:U487	324.035	306.025	Up
U488:U488	324.035	306.025	Up
U489:U489	324.035	306.025	Up
A490:U497	2603.379	2585.369	ANAGAAAUp
U498:U498	324.035	306.025	Up
U499:U499	324.035	306.025	Up
G500:U502	958.117	940.107	GNUp
A503:U504	653.087	635.077	AUp
G505:U508	1303.165	1285.155	GGNUp
G509:N511	901.179		GAN

According to the in silico digestion, there are 41 cleavage products with the length larger than 4 bases. The experimental result is the spectrum shown in Fig. 10.

There were 37 cleavage products identified in the spectrum. Among the 4 unidentified fragments, two were at the beginning of the sequence, which might not be transcribed, and the other two were of relatively large molecular weight: 5918 Da and 7708 Da.

The sequence coverage for a single reaction was usually not high (less than 70%), however for each fragment, there were in total 4 base-specific cleavage reactions. In Positive dU reaction (Fig. 11), 35 out of 37 cleavage products were identified, while in the two negative strand reactions, the coverage were 28 out of 30 (Fig. 12), and 36 out of 42 (Fig. 13). Data from these 4 reactions were combined together.

Figure 10. Shown above is the spectrum from WSN HA gene fragment 1, positive strand dC reaction. The sequence was first in silico digested using Mongo Oligo Mass Calculator v2.06. After cleavage, there were 41 cleavage products larger than 4 bases on the in silico digest list. After comparison, 37 peaks in this spectrum were identified as cleavage products. The missing peaks were either at the beginning of the sequence, or too large to get ionized.

Figure 11. Shown above is the spectrum from WSN HA gene fragment 1, positive strand dU reaction. Related parameters were set on Mongo Oligo Mass Calculator. After cleavage, 37 cleavage products were of interest on the in silico digest list. In this reaction, there were only 2 missing peaks. The rest 35 peaks were all found in the spectrum with relatively high intensity.

Figure 12. Shown above is the spectrum from WSN HA gene fragment 1, negative strand dC reaction. In this reaction, T7 promoter was added at the 5' end of reverse primer when PCR amplification was performed. As a result, the negative strand of the PCR product was transcribed. Similar calculation was done using Mongo Oligo Mass Calculator. There were 30 cleavage products with the length of more than 4 bases. The valuable peaks were less than the counterpart in the positive strand because in this direction, there were more rU in the sequenced, which means there were more cleavage site, resulting in the increasement of smaller fragments.

Figure 13. Shown above is the spectrum from WSN HA gene fragment 1, negative strand dU reaction. There were 42 cleavage products of interest on the in silico digest list. 36 of them were identified in the spectrum

Demonstrated by the above spectra, underlined areas were not covered in all the four reactions. For this 511 bases fragment, the sequence coverage is about 92%.

5'-CCAAAATGAAGGCAAGACTACTTGGTCCTGTTATGTGCACTTGCAGCTACAGATGC
AGACACAATATGTATAGGCTACCATGCGAACAACTCAACCGACACTGTTGACACACT
ACTCGAGAAGAATGTGACAGTGACACATTCTGTTAACCTGCTCGAAGACAGCCACAA
CGGGAACTATGTAAATTAAAAGGAATAGCCCCACTACAATTGGGGAAATGTAACAT
CACCGGATGGCTCTTGGGAAACCCAGAATGCGACTCACTGCTTCCAGCGAGATCATG
GTCCTACATTGTAGAAACACCAAACCTCTGAGAATGGAGCATGTTATCCAGGAGATTT
CATCGACTATGAGGAACTGAGGGAGCAATTGAGCTCAGTATCATCATTAGAAAGATT
CGAAATATTTCCCAAGGAAAGTTCATGGCCCAACCACAACACACTCAACGGAGTAAC
AGTATCATGCTCCCATAGGGGAAAAAGCAGTTTTTACAGAAATTTGCTATGGCTGAC
-3'

3.1.2 Fragment Two to WSN HA

The PCR product for the WSN HA gene fragment 2 is

```
5' -CAGAAATTTGCTATGGCTGACGAAGAAGGGGATTCATACCCAAAGCTGAACAA  
TTCTATGTGAACAATAAAGGGAAAGAAGTCCTTGTACTATGGGGTGTTTCATCACCC  
GTCTAGCAGTGATGAGCAACAGAGTCTCTATCGTAATGGAAATGCTTATGTCTCTGT  
AGCGTCTTCAAATTATAACAGGAGATTACCCCGGAAATAGCTGCAAGACCCAAAGT  
AAGAGATCAACATGGGAGGATGAACTATTACTGGACCTTGCTAGAACCCGGAGACAC  
AATAATATTTGAGGCAACTGGTAATCTAATAGCACCATGGTATGCTTTCGCACTGAG  
TAGAGGGTTTGGGTCCGGCATCATCACCTCAAACGCGTCAATGCATGAGTGTAACAC  
GAAGTGTC-3'
```

The experimental spectra from the 4 base-specific cleavage reactions were shown in Fig. 14, Fig. 15, Fig. 16 and Fig. 17.

In positive strand, dC reaction detected 31 peaks out of 40, dU reaction detected 35 peaks out of 42. In negative strand, dC reaction detected 32 peaks out of 39, while dU reaction detected 35 peaks out of 44.

Figure 14. Shown above is the spectrum from WSN HA gene fragment 2, positive strand dC reaction. There were 40 cleavage products of interest on the in silico digest list. 31 of them were identified in the spectrum. The peaks that are smaller than 4000 Da were labeled by their monoisotopic mass, and the relatively larger fragments like 5782 Da and 5919 Da in the figure may be mislabeled by 1 Da as the isotope patterns were not completely resolved.

Figure 15. Shown above is the spectrum from WSN HA gene fragment 2, positive strand dU reaction. There were 42 cleavage products of interest on the in silico digest list. 35 of them were identified in the spectrum. The largest fragment that can be identified in this spectrum is <UGAGUAGAGGGUUUGGUC> whose molecular weight is 6155.8 Da. The isotope patterns were not completely resolved.

Figure 16. Shown above is the spectrum from WSN HA gene fragment 2, negative strand dC reaction. There were 39 cleavage products of interest on the in silico digest list. 32 of them were identified in the spectrum. The mass peak of 4988 came from the fragment <GCCGGACCCAAACCCU>.

Figure 17. Shown above is the spectrum from WSN HA gene fragment 2, negative strand dU reaction. There were 44 cleavage products of interest on the in silico digest list. 35 of them were identified in the spectrum. In this spectrum, the mass peak 5363 for <UGUUAUAUUUGAAGAC> and the mass peak 5466 for fragment <GUUGAGGUGAUGC> were of good resolution and signal intensity. The largest peak we expected at 7057.9 Da was lost.

Peak patterns from the four spectra of fragment 2 were analyzed. The underlined areas in the sequence were not covered in all the four reactions. For this 404 bases fragment, the sequence coverage is about 94%.

5' - CAGAAATTTGCTATGGCTTGACGAAGAAGGGGGATTCATACCCAAAGCTGAACAA
TTCCTATGTGAACAATAAAGGGAAAGAAGTCCTTGTA CTATGGGGTGTTTCATCACCC
GTCTAGCAGTGATGAGCAACAGAGTCTCTATCGTAATGGAAATGCTTATGTCTCTGT
AGCGTCTTCAAATTATAACAGGAGATTCACCCCGGAAATAGCTGCAAGACCCAAAGT
AAGAGATCAACATGGGAGGATGAACTATTACTGGACCTTGCTAGAACCCGGAGACAC
AATAATATTTGAGGCAACTGGTAATCTAATAGCACCATGGTATGCTTTCGCACTGAG
TAGAGGGTTTGGGTCCGGCATCATCACCTCAAACGCGTCAATGCATGAGTGTAACAC
GAAGTGTC - 3'

3.1.3 Fragment Three to WSN HA

The PCR product for the WSN HA gene fragment 3 is

```
5' -TCCGGCATCATCACCTCAAACGCGTCAATGCATGAGTGTAACACGAAGTGTCAA
ACACCCAGGGAGCTATAAACAGCAGTCTCCCTTTCCAGAATATACACCCAGTCACA
ATAGGAGAGTGCCCAAATATGTCAGGAGTACCAAATTGAGGATGGTTACAGGACTA
AGAAACATCCCATCCATTCAATCCAGAGGTCTATTTGGAGCCATTGCTGGTTTTATT
GAGGGGGGATGGACTGGAATGATAGATGGATGGTATGGTTATCATCATCAGAATGAA
CAGGGATCAGGCTATGCAGCGGATCAAAAAGCACACAAAATGCCATTAACGGGATT
ACAAACAAGGTGAACTCTGTTATCGAGAAAATGAACACTCAATTCACAGCTGTGGGT
AAAGAATTCAACAACTTAGAAAAAAGGATGGAAAATTTAAATAAAAAAGTTGATGAT
GGATTTCTGGACATTTGGACATATAATGCAGAATTGTTAGTTCTACTGGAAAATGAA
AGGACTTTGGATTTCCATGACTTAAATGTGAAGAATCTGTACGAGAAAGTAAAAAGC
CAATTAAAGAATAATGCCAAAGAAATCGGAAATGG-3'
```

The experimental spectra from the 4 base-specific cleavage reactions were shown in Fig. 18, Fig. 19, Fig. 20 and Fig. 21.

In positive strand, dC reaction detected 49 peaks out of 53, dU reaction detected 41 peaks out of 44. In negative strand, dC reaction detected 30 peaks out of 35, while dU reaction detected 36 peaks out of 46.

Figure 18. Shown above is the spectrum from WSN HA gene fragment 3, positive strand dC reaction. There were 53 cleavage products of interest on the in silico digest list. 49 of them were identified in the spectrum. In this reaction, most of the fragments drop in the molecular weight range between 1000 Da ~ 4000 Da.

Figure 19. Shown above is the spectrum from WSN HA gene fragment 3, positive strand dU reaction. There were 44 cleavage products of interest on the *in silico* digest list. 41 of them were identified in the spectrum.

Figure 20. Shown above is the spectrum from WSN HA gene fragment 3, negative strand dC reaction. There were 35 cleavage products of interest on the in silico digest list. 30 of them were identified in the spectrum.

Figure 21. Shown above is the spectrum from WSN HA gene fragment 3, negative strand dU reaction. There were 46 cleavage products of interest on the in silico digest list. 36 of them were identified in the spectrum.

Peak patterns from the four spectra of fragment 3 were analyzed. The underlined areas in the sequence were not covered in all the four reactions. For this 602 bases fragment, the sequence coverage is about 91.2%.

5' - TCCGGCATCATCACCTCAAACGCGTCAATGCATGAGTGTAACACGAAGTGTCAA
ACACCCAGGGAGCTATAAACAGCAGTCTCCCTTTCCAGAATATACACCCAGTCACA
ATAGGAGAGTGCCCAAATATGTCAGGAGTACCAAATTGAGGATGGTTACAGGACTA
AGAAACATCCCATCCATTCAATCCAGAGGTCTATTTGGAGCCATTGCTGGTTTTATT
GAGGGGGGATGGACTGGAATGATAGATGGATGGTATGGTTATCATCATCAGAATGAA
CAGGGATCAGGCTATGCAGCGGATCAAAAAAGCACACAAAATGCCATTAACGGGATT
ACAAACAAGGTGAACTCTGTTATCGAGAAAATGAACACTCAATTCACAGCTGTGGGT
AAAGAATTCAACAACCTTAGAAAAAAGGATGGAAAATTTAAATAAAAAAGTTGATGAT
GGATTTCTGGACATTTGGACATATAATGCAGAATTGTTAGTTCTACTGGAAAATGAA
AGGACTTTGGATTTCCATGACTTAAATGTGAAGAATCTGTACGAGAAAGTAAAAAGC
CAATTAAAGAATAATGCCAAAGAAATCGGAAATGG-3'

3.1.4 Fragment Four to WSN HA

The PCR product for the WSN HA gene fragment 4 is

```
5' - TTTCTGGACATTTGGACATATAATGCAGAATTGTTAGTTCTACTGGAAAATGAA  
AGGACTTTGGATTTCCATGACTTAAATGTGAAGAATCTGTACGAGAAAGTAAAAAGC  
CAATTAAAGAATAATGCCAAAGAAATCGGAAATGGGTGTTTTGAGTTCTACCACAAG  
TGTGACAATGAATGCATGGAAAGTGTAAGAAATGGGACTTATGATTATCCAAAATAT  
TCAGAGGAATCAAAGTTGAACAGGGAAAAGATAGATGGAGTGAAATTGGAATCAATG  
GGGGTGTATCAGATTCTGGCGATCTACTCAACTGTCGCCAGTTCACTGGTGCTTTTG  
GTCTCCCTGGGGGCAATCAGTTTCTGGATGTGTTCTAATGG-3'
```

The experimental spectra from the 4 base-specific cleavage reactions were shown in Fig. 22, Fig. 23, Fig. 24 and Fig. 25.

In positive strand, dC reaction detected all the 29 peaks, dU reaction detected 22 peaks out of 26. In negative strand, dC reaction detected 22 peaks out of 24, while dU reaction detected 33 peaks out of 36.

Figure 22 Shown above is the spectrum from WSN HA gene fragment 4, positive strand dC reaction. There were 29 cleavage products of interest on the in silico digest list and all of them were identified in the spectrum.

Figure 23 Shown above is the spectrum from WSN HA gene fragment 4, positive strand dU reaction. There were 26 cleavage products of interest on the in silico digest list. 22 of them were identified in the spectrum.

Figure 24 Shown above is the spectrum from WSN HA gene fragment 4, negative strand dC reaction. There were 24 cleavage products of interest on the in silico digest list. 22 of them were identified in the spectrum.

Figure 25 Shown above is the spectrum from WSN HA gene fragment 4, negative strand dU reaction. There were 36 cleavage products of interest on the in silico digest list. 33 of them were identified in the spectrum.

Peak patterns from the four spectra of fragment 4 were analyzed. The underlined areas in the sequence were not covered in all the four reactions. For this 382 bases fragment, the sequence coverage is about 97.6%.

```
5' - TTTCTGGACATTTGGACATATAATGCAGAATTGTTAGTTCTACTGGAAAATGAA
AGGACTTTGGATTTCCATGACTTAAATGTGAAGAATCTGTACGAGAAAAGTAAAAAGC
CAATTAAAGAATAATGCCAAAGAAATCGGAAATGGGTGTTTTGAGTTCTACCACAAG
TGTGACAATGAATGCATGGAAAGTGTAAGAAATGGGACTTATGATTATCCAAAATAT
TCAGAGGAATCAAAGTTGAACAGGGAAAAGATAGATGGAGTGAAATTGGAATCAATG
GGGGTGATCAGATTCTGGCGATCTACTCAACTGTCGCCAGTTCACTGGTGCTTTTG
GTCTCCCTGGGGGCAATCAGTTTCTGGATGTGTTCTAATGG-3'
```

3.1.5 Coverage of HA gene

Results from the four fragments (Table 2) were combined to get the overall coverage of comparative sequencing for WSN HA gene: 1563 out of 1676 bases were covered in the assay and the overall coverage is 93.3%

Table 2. *In silico* and experimental peak numbers of WSN HA gene

Reaction	<i>in silico</i> peak No.	Experimental peak No.
+1C	41	37
+1U	37	35
+2C	40	31
+2U	42	35
+3C	53	49
+3U	44	41
+4C	29	29
+4U	26	22
-1C	30	28
-1U	42	36
-2C	39	32
-2U	44	35
-3C	35	30
-3U	46	36
-4C	24	22
-4U	36	33

The primers were designed in the conserved regions between WSN and PR8 viral strains, so the same approach could be performed directly to the HA genes of PR8 viral strain. (Results were not shown here)

3.2 Analysis of HA genes from WSN/PR8 mixture

The different sequences of HA gene between WSN and PR8 could result in different peak patterns of a certain base-specific cleavage reaction. For example, in fragment 3, positive strand dU reaction (Fig. 26), WSN has 9 unique peaks: 1671 Da, 2361 Da, 2603 Da, 2635 Da, 2925 Da, 3082 Da, 3183 Da, 4954 Da, and 5111 Da. These peaks were not in the list of counterpart reaction of PR8 viral strain. At the same time, PR8 also has 6 unique peaks that may not appear in WSN reaction: 2258 Da, 2322 Da, 2502 Da, 2619 Da, 2651 Da, 4531 Da.

These unique peaks could be used as fingerprint to test the sensitivity of the base-specific cleavage assay.

Figure 26. The Venn Diagram of the number of peaks in HA fragment 3 positive dU reaction. The two viral strains had 35 cleavage products in common. WSN had 9 unique peaks and PR8 had 6 unique peaks. These unique peaks not only could be used to detect each viral strain, but also can serve as markers to test the sensitivity of the assay by mixing WSN and PR8 cDNA in different ratios.

Viral cDNA of WSN and PR8 was mixed in five different ratios: WSN to PR8 at (1) 1:5; (2) 1:3; (3) 1:1; (4) 3:1; and (5) 5:1.

In 1:5 and 5:1 ratio, only the majority sequence is observed. Fig. 27 showed the spectrum of cleavage product from the reaction which used cDNA of WSN and PR8 in 5:1 ratio. Seven out nine unique peaks for WSN were detected but none of PR8 unique peaks was found.

Figure 27. Shown in this spectrum is mixture of WSN to PR8 cDNA in 5:1 ratio. Seven of WSN unique peaks were detected: 1671, 2361, 2603, 2635, 3082, 3183, 4954. None of PR8 unique peaks was found in the spectrum. The result is almost the same as pure WSN.

Figure 28. Shown in this spectrum is mixture of WSN to PR8 cDNA in 1:3 ratio. Two of WSN unique peaks were detected: 1671, 2361, while 5 of PR8 unique peaks was found: 2258, 2322, 2651, 2502, 2619.

In 1:3 and 3:1 ratio, both of WSN and PR8 could be detected. More unique peaks could be detected from the major material. Fig. 28 showed the spectrum of cleavage product from the reaction which used cDNA of WSN and PR8 in a 1:3 ratio. Two unique peaks for WSN and five unique peaks for PR8 were detected in this spectrum.

When WSN and PR8 were mixed in 1:1 ratio, there was a significant decrease of intensity from those unique peaks. This might be caused by the charge competition effect. According to these results, base-specific cleavage assay has the ability to detect virus sample in a mixture, but the sensitivity need to be further improved.

3.3 Analysis of NA genes from WSN & PR8

NA gene is about 1400 bases in length. Three pairs of primers (Table 3) were designed to amplify 3 fragments to cover the full length of NA gene (Fig. 29).

Figure 29. Shown above is the design of primers for NA gene. The NA gene is about 1400 bases in length, and the primers were all designed in the conserved region between WSN and PR8 viral strains. Three PCR reactions were performed in order to cover the full length of the NA gene. The adjacent PCR products would have an overlapped region.

Table 3. Primers for NA genes All the “F” primer had a T7-promoter sequence at the 5’ end. “+” reaction would transcribe the sense strand sequence while “-” reaction would transcribe the antisense strand sequence.

Primer name	Primer sequence
NA+1F	CAGTAATACGACTCACTATAGGGAGACAGAAAAATAATAACCATG
NA+1R	CCGATTGTTAGCCAGCCCATGCC
NA-1F	CAGTAATACGACTCACTATAGGGAGATTGTTAGCCAGCCCATGCC
NA-1R	GAATCCAAACCAGAAAAATAATAACCATG
NA+2F	CAGTAATACGACTCACTATAGGGAGATTGGTCAGCAAGTGCATG
NA+2R	GTCCCTTCCTATCCAAACACCATTAAC
NA-2F	CAGTAATACGACTCACTATAGGGAGAGTCCTTCCTATCCAAACAC
NA-2R	GGTCAGCAAGTGCAATGTCATGATGGC
NA+3F	CAGTAATACGACTCACTATAGGGAGATGGAGCAACGGAGTAAAG
NA+3R	GGCAACTCAGCACCGTCTGGCC
NA-3F	CAGTAATACGACTCACTATAGGGAGAGGCAACTCAGCACCGTCTGGC
NA-3R	GATGGAGCAACGGAGTAAAGGG

3.3.1 Fragment One to WSN NA

The PCR product for the WSN NA gene fragment 1 is

```
5' -CAGAAAATAATAACCATTGGGTCAATCTGTATGGTAGTCGGAATAATTAGCCTA
ATATTGCAAATAGGAAATATAATCTCAATATGGATTAGCCATTCAATTCAAACCGGA
AATCAAAACCATACTGGAATATGCAACCAAGGCAGCATTACCTATAAAGTTGTTGCT
GGGCAGGACTCAACTTCAGTGATATTAACCGGCAATTCATCTCTTTGTCCCATCCGT
GGGTGGGCTATACACAGCAAAGACAATGGCATAAGAATTGGTTCCAAAGGAGACGTT
TTTGTTCATAAGAGAGCCTTTTATTTTCATGTTCTCACTTGGAATGCAGGACCTTTTTT
CTGACTCAAGGCGCCTTACTGAATGACAAGCATTCAAATGGGACCGTTAAGGACAGA
AGCCCTTATAGGGCCTTAATGAGCTGCCCTGTCGGTGAAGCTCCGTCCCCGTACAAT
TCAAGGTTTGAATCGGTTGCTTGGTCAGCAAGTGCATGTCATGATGGCATGGGCTGG
CTAACAATCGG-3'
```

The experimental spectra from the 4 base-specific cleavage reactions were shown in Fig. 30, Fig. 31, Fig. 32 and Fig. 33.

In positive strand, dC reaction detected 30 peaks out of 40, dU reaction detected 39 peaks out of 48. In negative strand, dC reaction detected 27 peaks out of 35, while dU reaction detected 36 peaks out of 42.

Figure 30. Shown above is the spectrum from WSN NA gene fragment 1, positive strand dC reaction. There were 40 cleavage products of interest on the in silico digest list. 30 of them were identified in the spectrum.

Figure 31. Shown above is the spectrum from WSN NA gene fragment 1, positive strand dU reaction. There were 48 cleavage products of interest on the in silico digest list. 39 of them were identified in the spectrum.

Figure 32. Shown above is the spectrum from WSN NA gene fragment 1, negative strand dC reaction. There were 35 cleavage products of interest on the in silico digest list. 27 of them were identified in the spectrum.

Figure 33. Shown above is the spectrum from WSN NA gene fragment 1, negative strand dU reaction. There were 42 cleavage products of interest on the in silico digest list. 36 of them were identified in the spectrum.

3.3.2 Fragment Two to WSN NA

The PCR product for the WSN NA gene fragment 2 is

```
5' -TTGGTCAGCAAGTGCATGTCATGATGGCATGGGCTGGCTAACAATCGGAATTC
TGGTCCAGATAATGGAGCAGTGGCTGTATTAAAATACAACGGCATAATAACTGAAAC
CATAAAAAGTTGGAGGAAGAAAATATTGAGAACACAAGAGTCTGAATGTACCTGTGT
AAATGGTTCATGTTTTACCATAATGACCGATGGCCCAAGTAATGGGCTGGCCTCGTA
CAAAATTTTCAAGATCGAGAAGGGGAAGGTTACTAAATCAATAGAGTTGAATGCACC
TAATTCTCACTACGAGGAATGTTCTGTACCCTGATACCGGCAAAGTGATGTGTGT
GTGCAGAGACAATTGGCACGGTTCGAACCGACCATGGGTGTCCTTCGACCAAAACCT
AGATTATCAAATAGGATACATCTGCAGTGGGGTTTTTCGGTGACAACCCGCGTCCCAA
AGATGGAACAGGCAGCTGTGGCCCAGTGTCTGCTGATGGAGCAAACGGAGTAAAGGG
ATTTTCATATAGGTATGGTAATGGTGTGTTGGATAGGAAGGAC - 3'
```

The experimental spectra from the 4 base-specific cleavage reactions were shown in Fig. 34, Fig. 35, Fig. 36 and Fig. 37.

In positive strand, dC reaction detected 30 peaks out of 40, dU reaction detected 34 peaks out of 45. In negative strand, dC reaction detected all the 35 peaks, while dU reaction detected 32 peaks out of 40.

Figure 34. Shown above is the spectrum from WSN NA gene fragment 2, positive strand dC reaction. There were 40 cleavage products of interest on the in silico digest list. 30 of them were identified in the spectrum.

Figure 35. Shown above is the spectrum from WSN NA gene fragment 2, positive strand dU reaction. There were 45 cleavage products of interest on the in silico digest list. 34 of them were identified in the spectrum.

Figure 36. Shown above is the spectrum from WSN NA gene fragment 2, negative strand dC reaction. There were 35 cleavage products of interest on the in silico digest list and all of them were identified in the spectrum.

Figure 37. Shown above is the spectrum from WSN NA gene fragment 2, negative strand dU reaction. There were 40 cleavage products of interest on the in silico digest list. 32 of them were identified in the spectrum.

3.3.3 Fragment Three to WSN NA

The PCR product for the WSN NA gene fragment 3 is

```
5' -TGGAGCAAACGGAGTAAAGGGATTTTCATATAGGTATGGTAATGGTGTGTTGGAT
AGGAAGGACTAAAAGTGACAGTTCCAGACATGGGTTTGAGATGATTTGGGATCCTAA
TGGATGGACAGAGACTGATAGTAGGTTCTCTGTGAGACAAGATGTTGTGGCAATGAC
TGATCGGTCAGGGTACAGCGGAAGTTTCGTTCAACATCCTGAGCTAACAGGGCTAGA
CTGTATGAGGCCTTGCTTCTGGGTTGAATTAATCAGGGGGCTACCTGAGGAGGACGC
AATCTGGACTAGTGGGAGCATCATTTCTTTTTGTGGTGTGAATGGTGATACTGTAGA
TTGGTCTTGCCAGACGGTGCTGAGTTGCC-3'
```

The experimental spectra from the 4 base-specific cleavage reactions were shown in Fig. 38, Fig. 39, Fig. 40 and Fig. 41.

In positive strand, dC reaction detected 22 peaks out of 23, dU reaction detected 20 peaks out of 30. In negative strand, dC reaction detected 15 peaks out of 27, while dU reaction detected 19 peaks out of 25.

Figure 38. Shown above is the spectrum from WSN NA gene fragment 3, positive strand dC reaction. There were 23 cleavage products of interest on the in silico digest list. 22 of them were identified in the spectrum.

Figure 39. Shown above is the spectrum from WSN NA gene fragment 3, positive strand dU reaction. There were 30 cleavage products of interest on the in silico digest list. 20 of them were identified in the spectrum.

Figure 40. Shown above is the spectrum from WSN NA gene fragment 3, negative strand dC reaction. There were 27 cleavage products of interest on the in silico digest list. 15 of them were identified in the spectrum.

Figure 41. Shown above is the spectrum from WSN NA gene fragment 3, negative strand dU reaction. There were 25 cleavage products of interest on the in silico digest list. 19 of them were identified in the spectrum.

Results from the three fragments (Table 4) were combined to get the overall coverage of comparative sequencing for WSN NA gene. The overall sequence coverage was 90%

Table 4. *In silico* and experimental peak numbers of WSN NA gene

Reaction	<i>in silico</i> peak num	Experimental peak num
+1C	40	30
+1U	48	39
+2C	40	30
+2U	45	34
+3C	23	22
+3U	30	20
-1C	35	27
-1U	42	36
-2C	35	35
-2U	40	32
-3C	27	15
-3U	25	19

The same process was performed to PR8 NA gene (Table 5) and the overall sequence coverage was 94%. (Spectra were not shown here)

Table 5. *In silico* and experimental peak numbers of PR8 NA gene

Reaction	<i>in silico</i> peak num	Experimental peak num
+1C	41	39
+1U	47	42
+2C	40	36
+2U	44	37
+3C	23	21
+3U	31	22
-1C	33	28
-1U	43	34
-2C	39	26
-2U	38	31
-3C	27	24
-3U	24	16

Chapter Four. Surveillance of genetic drift

4.1 Validation of Known Sequence Changes

Random genetic drift occurs through a succession of point mutations in the Influenza HA or NA genes, resulting in an accumulation of amino acid residue changes that attenuate the host's immunity. This process occurs in response to the selective pressure brought about by the buildup in the human population of immunity to the extant viral strains. HA varies in this manner by an average of 3.5 accepted amino acid changes per year.

Base-specific cleavage & MALDI-MS assay has potential for application to the surveillance of these genetic drifts. When a point mutation occurs, up to 10 “observations” could be found in the mass spectra from the four base-specific cleavage reactions. In real cases, it is not necessary to see all these 10 observations to confirm a point mutation, because some observation might be a fragment with too long or too short size, or be overlapped by another existing peak. Most of the time, two observations are enough to find a base change in the sequence.

After RT-PCR, the cDNA of WSN viral strain was amplified with the universal primers:

Bm-HA-1: 5'-TATTCGTCTCAGGGAGCAAAAGCAGGGG

Bm-NS-890R: ATATCGTCTCGTATTAGTAGAAACAAGGGTGTTTT

The amplified HA gene was sequenced and the result showed that there were 13 sites different from the sequence in NCBI online database (Table 6).

Table6. Comparison of sequencing result with the NCBI database (CY009604)

Position	Sequenced result	Database
242	A	G
375	G	A
460	C	A
471	T	C
485	A	G
529	G	A
626	G	A
648	G	A
654	G	T
675	C	T
741	A	C
850	G	A
1543	G	A

For example at position 460, in the positive strand, dC reaction, the cleavage product containing site 460 should be

⁴⁵⁷ CAAAGGAGU ⁴⁶⁵ molecular weight 2964.4 Da or

⁴⁵⁷ CAACGGAGU ⁴⁶⁵ molecular weight 2924.4 Da

The absence of the 2965 Da peak and the presence of the 2925 Da peak confirmed the correct sequence to be : ⁴⁵⁷ CAACGGAGU ⁴⁶⁵ (Fig. 42)

In the positive strand, dU reaction, the cleavage product containing site 460 would be:

⁴⁵⁸ AAAGGAGTAAC ⁴⁶⁸ molecular weight 3636.5 Da or

⁴⁵⁸ AAC ⁴⁶⁰ ⁴⁶¹ GGAGTAAC ⁴⁶⁸ molecular weight 981.2 Da and 2649.4 Da

The absence of the 3637 Da peak (Fig. 43) and the presence of the 2650 Da peak (Fig. 44) conformed the correct sequence to be : ⁴⁵⁸ AAC ⁴⁶⁰ ⁴⁶¹GGAGTAAC ⁴⁶⁸. The spectra clearly show that there is a cytosine instead of an adenosine at site 460

These known sequence differences between the viral genome and the NCBI database listed in Table 6 can be validated in the same way by comparing the related mass peak changes. This approach showed convincingly that point mutations can be detected in multiple different spectra. The advantage of multiple independent verifications is unique in our approach, and not available in direct sequencing.

Figure 42 dC reaction of the positive strand. Shown in the spectrum is m/z 2910-2990 Da range. There should be a peak around 2925 Da if site 460 is “C” or a peak around 2965 Da if site 460 is “A”

Figure 43 dU reaction of the positive strand. Shown in the spectrum is m/z 3560-3720 Da range. There should be a peak around 3637 Da if site 460 is “A”

Figure 44 dU reaction of the positive strand. Shown in the spectrum is m/z 2640-2670 Da range. There should be peak around 2650 Da if site 460 is “C”.

4.2 Monitoring of Unknown Sequence Changes

We can manually analyzed the experimental spectra to validate the sequence changes by comparing with the in silico digest result only if those sequence changes are already known. If the sequence changes are unknown, the experimental peak list could be extracted from the spectrum and analysed with in silico digest peak list using certain software(118).

We collaborated with Dr. Chen Xin's group from Division of Mathematical Sciences, SPMS of NTU in optimizing the automatic analysis method of our experimental data.

WSN HA fragment 2 was used as an example. The experimental mass peak list was extracted from the spectra of four base-specific cleavage reactions. The extracted mass peak patterns were searched against the in silico digest list of the reference sequence from NCBI website(Fig. 45). The raw data could be automatically calibrated and the error tolerance was normally set as 0.3 Da. All the "observations" of new emergent peaks, missing peaks or shifted peaks were used for speculation of possible sequence changes.

```

0  CAGTAATACGACTCACTATAGGGAGATTTCGAAATATTTCCCAAGGAAAGTTCATGGCCCA
   ||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| |||||
   CAGcAATACGAtTCACTATAGGGAGATTTCGAAATATTTCCCAAGGAAAGTTCATaGCCCA
       o       e                               r
60  ACCACAACACACTCAAAGGAGTAACAGCATCATGCTCCCATGGGGGAAAAAGCAGTTTTT
   ||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| |||||
   ACCACAACACACTCAAtGGAGTAACAGtATCATGCTCCCATaGGGGcAAAAGCAGTTTTT
               q       a       c       p
120  ACAGAAATTTGCTATGGCTGACGAAAAAGGGGATTTCATACCCAAAGCTGAACAATTCCT
   ||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| |||||
   ACAGAAATTTGCTATGGCTGAtGAAAcAGGGGATTTCATACCCAAAGCTGAACAATTCCT
                   k       g
180  ATGTGAACAATAAAGGAAAGAAGTCCTTGTACTATGGGGTGTTCATCACCCGTCTAGCA
   ||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| |||||
   ATGTGAACAATAAAGGtAAAtAAGTCCTTGTACTATGGGGTGTTCATCACCCGTaTAGCA
                   i       l                               f
240  GTAATGAGCAACAGAGTCTCTATCATAATGTAAATGCTTATGTCTCTGTAGTGTCTTCAA
   ||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| |||||
   GTAATGgGCAACAGAGTCTCTATCgTAAgGTAAATGCTTATGTCTCTGTAGcGTCTTCAA
                   h               n       m                               d
300  ATTATAACAGGAGATTACCCCGGAAATAGCTGCAAGACCCAAAGTAAGAGATCAACCTG
   ||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| |||||
   ATTATAACAGGAGATTACCCCGGAAATAGCTGCAAGACCCAAAGTAAGAGATaAACCTG
                                           j
360  GGAGGATGAACTATTACTGGACCTTGCTAGAACCCGGAGACACAATAATATTGAGGCAA
   ||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| |||||
   GGAGGATGAACTATTACTGGACCTTGCTAGAACCCGGAGACACAATAATATTGAGGCAA

420  CTGGTAATCTAATAGCACCATGGTATGCTTTCGCACTGAGTAGAGGATTGGGTCCGGCA
   ||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| |||||
   CTGGTAATCTAATAGCACCATGGTATGCTTTCGCACTGAGTgGAGGATTGGGTCCGGCA
                                   b
480  TCATCACCTCAAACGCGTCAATGCATGAGTGTAACACGAAGTGCAAACACCCAGG
   ||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| |||||
   TCATCACCTCAAACGCGTCAATGCATGAGTGTAACACGAAGTGCAAACACCCAGG

```

Figure 45. Output of automatic analysis. The sequence was from the second segment of HA gene of WSN viral strain. Molecular weight information was exported from MALDI mass spectra of base-specific cleavage products. Experimental data from four reactions were combined in the software and searched against reference sequence and totally 18 possible sequence changes were found.

Table 7. Result of automatic analysis. 18 possible sequence changes were found.

No.	Position	Experiment	Reference	Score
a	88	T	C	302
b	462	G	A	151
c	102	A	G	147
d	292	C	T	124
e	12	T	C	107
f	235	A	C	104
g	147	C	A	97
h	247	G	A	81
i	197	T	G	76
j	354	A	C	76
k	142	T	C	34
l	201	T	G	34
m	269	G	T	32
n	265	G	A	48
o	4	C	T	30
p	107	C	A	30
q	77	T	A	26
r	55	A	G	21

Totally 18 possible sequence changes were found by comparing the experimental data with the reference sequences (Table 7). For surveillance of influenza genes, the sequence changes were normally limited to substitutions. As a result, insertions and deletions were not included to avoid large amount of false positive results. For example, result a: (88, s, t, 302) showed one sequence change happened at position “88”, the type of the sequence change was “substitution”, according to the experimental data, the base at that position should be “t”, and the signal to noise ratio of the relevant peak was 302.

Among all the 18 possible sequence changes, two dropped in the T7 promoter sequence (“o” and “e”). Four of them identified the correct sequence change, but could only locate the sequence change in a certain area (“h” “m” “j” “b”). Two of them found uncertain sequence changes at the accurate positions(“q” “g”). The above 6 sequence were partially

identified because only one “observation” was available from all 4 reactions. The other 4 identified the correct sequence changes at the accurate positions (“a” “c” “n” “d”). These sequence changes were identified based on at least 2 “observations”. The rest 6 were false positive result (“r” “p” “k” “i” “l” “f”).

4.3 Surveillance of genetic drift

Sequence changes in HA genes and NA genes will possibly result in resistance of influenza virus to the neuraminidase inhibitor drugs, such as oseltamivir (Tamiflu), zanamivir (Relenza).

One viral sample WSN-(T5A) from influenza A /WSN/33 (H1N1) strain that was suspected to have some unknown mutations in HA gene was used for evaluation of the base-specific cleavage & MALDI-MS assay.

Base-specific cleavage & MALDI-MS assay was performed to analyse the HA gene of WSN-(T5A) sample and all spectra from the experiment were compared with the counterpart from in silico digest list.

In total, three significant “observations” were found, one was from fragment 3 negative strand dC reaction and the other two were from fragment 4 positive strand dC and dU reactions.

In fragment 3 negative strand dC reaction, there is a significant mass peak shift from $[M+H]^+$ of 4346 Da to 4330 Da (Fig. 46).

This peak is from the fragment <GGAAAGGGAGAdCU>.

Table 8 . MW of dNTP and rNTP used in the assay

Base	Molecular Weight (Da)
dC	289.05
rU	306.03
rG	345.05
rA	329.05
rC	305.05
dU	290.03

In the cleavage product from the dC reaction, a minus 16 Da mass shift could most probably come from a G/A mutation (Table 8).

$$345.05 \text{ Da} - 329.05 \text{ Da} = 16 \text{ Da}$$

(A/C mutation was 40 Da mass shift while C/G mutation was 56Da)

In fragment 4 positive strand dC reaction, a mass peak at 2925 Da was missing(Fig. 47). The peak is from the fragment <dCAGdCGGAAU>. In the dU reaction, a mass peak at 1962 Da which is from the fragment <GGAAAdUC> was also missing(Fig. 48). These two missing peaks were at the same region and their disappearances were unlikely caused by charge competition effect. Thus we can draw a conclusion that there was at least one mutation in the sequence.

The cDNA of WSN-(T5A) was then sent for sequencing and the results agreed with the base-specific cleavage & MALDI MS assay. There is a C/T point mutation at site 945, resulting in an G/A change in the fragment 3 negative strand reaction. The relative cleavage product for sample WSN-(T5A) should be <GGAAAGGGAGAdCU> instead of <GGGAAGGGAGAdCU> for reference sequence. Another C/A mutation was found at site 1552, resulting in a mass shift from 2925 Da to 2965 Da in fragment 4 positive strand dC reaction and a peak missing at 1962 Da in the dU reaction. As the 2965 Da peak was also in the *in silico* digest list, it was possible to find the region where the mutation occurred but it was hard to identify it in this case.

Figure 47. WSN-(T5A) sample, fragment 4, dC reaction of the positive strand. The mass peak of 2925 Da from the fragment < dCAGdCGGAU > was missing.

Figure 48. WSN-(T5A) sample, fragment 4, dU reaction of the positive strand. The mass peak of 1962 Da from the fragment < GGAAAdUC > was missing.

Chapter Five. Data Analysis & Optimization

5.1 Evaluation of the MS data

The most valuable result from Base-specific cleavage & MALDI-MS assay is the MALDI spectra. One good spectrum need to meet two major requirements: (1) the spectrum should be of high quality; (2) the peaks in the spectrum should match well with the *in silico* digest result.(Fig. 49)

Figure 49. Spectrum Quality vs Match. These are the two basic requirements for a good spectrum from base-specific cleavage & MALDI-MS assay, quality and match. The spectral quality includes signal intensity, resolution, signal to noise ratio, *etc.* A spectrum with high quality is of great importance in data extraction step. A good spectrum also need to be well matched with the *in silico* digest result from the reference sequence.

In order to achieve high signal intensity of nucleic acid samples with 3-HPA as the matrix, it is necessary to use higher laser power, which will decrease the resolution of the mass peaks. It is always challenging to find a balance between signal intensity and the resolution.

In a good spectrum (Fig. 50), the mass peaks with molecular weight less than 4000 Da should be of high resolution and those mass peaks larger than 4000 Da could still be detected although not completely resolved. At the same time, the mass peak pattern should agree with the in silico digest result, which means most of the in silico digest fragment could be identified in this experimental spectrum.

In some of the spectra (Fig. 51), the signal intensity and resolution were both satisfactory, but after comparison with in silico digest list, a number of peaks with molecular weight less than 4000 Da were missing. This poor match spectrum might result in low sequence coverage of one reaction. The most probable reason for these spectra comes from the RNase A digest and desalting steps. Incomplete digestion or mis-cleavage of the RNase A and improper use of desalting resin could cause peak missing.

Due to charge competition effect, some large peaks were difficult to get ionized in certain reaction (Fig. 52) . In the spectra from these reactions, the sequence coverage mainly depends on the number of peaks in lower mass range.

Figure 50. Shown above is an example spectrum with high quality and perfect match. The spectrum was from the WSN viral strain, NA gene, fragment 2, negative strand dC reaction. The peaks in low mass range were all completely resolved and the high mass peak around 5000 Da still could be detected. There were 35 fragments from the in silico digest result, and all of them could be identified in the spectrum.

Figure 51. Shown above is an example spectrum with high quality and poor match. The spectrum was from the PR8 viral strain, NA gene, fragment 3, negative strand dU reaction. All the peaks detected were of high resolution and signal intensity. However, some of the important peaks were missing in the spectrum. There were 24 fragments in the in silico digest list, but only 16 of them were found.

Figure 52. Shown above is an example spectrum with significant charge competition. The spectrum was from the PR8 viral strain, NA gene, fragment 2, negative strand dC reaction. There were 39 fragments in the in silico digest list, and 26 of them were identified. The sequence coverage of this reaction is not low because quite a number of peaks were smaller than 3000 Da. All the missing peaks dropped in the high mass range(>4000 Da). The most probable cause of the missing peak is charge competition effect.

5.2 Cyclic Phosphate

Figure 53. Shown above is the forming of 2',3'-cyclic phosphodiester intermediate through RNase cleavage.

RNaseA cleavage reaction is believed to accomplish in two steps. The first step would generate a 2',3'-cyclic phosphodiester intermediate (Fig. 53), which is hydrolyzed to a 3'-monophosphate group in the second step.(179)

Unlike RNase T1 cleavage reaction, which would keep a portion of phosphodiester intermediate as cleavage product, RNase A is known to have the ability to drive the cleavage reaction to completion.

Using WSN viral strain, in Fragment 3, negative strand dC reaction, the molecular weight for cleavage products <dCAdCAU> , <dCAAdCU> and <AdCdCAU> are all 1561 Da, however, a significant peak at molecular weight of 1543 (Fig. 54), which refers to the 2',3'-cyclic phosphodiester intermediate of these fragments can be observed in the spectrum. Another peak in the same mass range at 1559 Da could be 2',3'-cyclic phosphodiester intermediate of the fragment <AGdCdCU> and <dCdCAGU>.

The peak at molecular weight of 1872 Da (Fig. 55) was actually 2',3'-cyclic phosphodiester intermediate from the fragment <dCdCAAAU> and <AAAdCdCAU> and the nearby peak 1888 Da was from the 2',3'-cyclic phosphodiester intermediate of fragment <AGAdCdCU>

Using WSN viral strain, in Fragment 3, negative strand dU reaction, some other 2',3'-cyclic phosphodiester intermediate were identified.

The peak at molecular weight of 2217 Da (Fig. 56) was 2',3'-cyclic phosphodiester intermediate from the fragment <AAAGdUdUC>.

The peaks at molecular weight of 2602 Da (Fig. 57) was 2',3'-cyclic phosphodiester intermediate from the fragment <AGdUAGAAC> . The peaks at molecular weight of 1968 Da (Fig. 58) was 2',3'-cyclic phosphodiester intermediate from the fragment <AGAAAC>.

In the spectra from other reactions, we can also find extra peaks compared with the in silico digest list. Some of them were sodium or potassium addition on the cleavage products,

but quite a number of them could be 2',3'-cyclic phosphodiester intermediate. These 2',3'-cyclic phosphodiester intermediate peaks sometimes overlap with normal cleavage products.

For example, 2',3'-cyclic phosphodiester intermediate for fragment <AGAdCdCU> was 1888 Da, while in the same reaction, the fragment <dCdCAAAU> was 1890 Da. The mass difference is only 2 Da. The isotopic patterns of these two products overlapped with each other. Without high resolution MALDI-MS, it was impossible to differentiate the two products and was quite likely to get a false positive result.

Figure 54. The peak at molecular weight of 1543 Da was of high signal intensity and high resolution. However, it could not be explained using any of the cleavage product in the in silico digest list. This peak could only be the 2',3'-cyclic phosphodiester intermediate from the fragment <dCA₂dCAU>. The adjacent peak 1559 Da was the 2',3'-cyclic phosphodiester intermediate from the fragment <AGdCdCU>

Figure 55. The peak 1872 Da was the 2',3'-cyclic phosphodiester intermediate from the fragment <dCdCAAAU> and the peak 1888 Da was the 2',3'-cyclic phosphodiester intermediate from the fragment <AGAdCdCU>

Figure 56. The peak 2217 Da was the 2',3'-cyclic phosphodiester intermediate from the fragment <AAAGdUdUC>

Figure 57. The peak 2602 Da was the 2',3'-cyclic phosphodiester intermediate from the fragment <AGdUAGAAC>

Figure 58. The peak 1968 Da was the 2',3'-cyclic phosphodiester intermediate from the fragment <AGAAAC>

5.3 Optimization of Matrix

The comparative sequencing approach using base-specific cleavage & MALDI-TOF MS assay can achieve high coverage by combining multiple reactions for a certain sequence. However, single reactions were rarely perfect matched with the *in silico* digest result. Despite those cleavage products that were too long (> 6000 Da) to get ionized, or too short (< 1400 Da) to be unique in a 500 bases sequence, some peaks which should appear did miss in the spectrum.

Although the reason for those missing peak is still under investigation, we did try to increase the spectrum quality by optimizing the matrix. Matrices were the essential part for MALDI-TOF MS and several different matrices had been reported suitable for nucleic acid analysis, such as 3-HPA, ATT, PA and THAP. We tried all these matrices separately for the cleavage product from our assay, and found that 3-HPA gave the most satisfactory performance.

After that, two of different matrices were mixed together and it is interesting that when 3-HPA and ATT were mixed in 1:1 ratio as a compound matrix, the signal to noise ratio and the resolution of the spectrum increased (Fig. 59).

Figure 59. MS spectra of PR8, fragment 3, positive strand dC reaction from 1500 to 4000 m/z range. A: 3-HPA matrix, B: ATT matrix, C: compound matrix of 3HPA and ATT. In the spectrum C, the signal to noise ratio and resolution for each peak was higher than those in spectrum A and B, which used 3-HPA and ATT only.

Under microscopic view of the matrix crystal, we found that the compound matrix was more evenly distributed than 3-HPA, which normally formed a “donut” as a result of its high surface tension (Fig. 60).

Figure 60. Microscopic view of matrix crystals. Images on the left were taken at 3X magnification. Images on the right were taken at 10X magnification of the boxed area. A: 3-HPA matrix, B: ATT matrix, C: compound matrix

According to our experiment, the compound matrix mixed from 3-HPA and ATT had a better performance than single matrix. The signal to noise ratio and resolution were increased while the quality of spectra was more consistent.

Chapter Six. Conclusion

In this work, we applied a comparative sequencing method based on high-throughput MALDI-MS for the characterization of two influenza A H1N1 viral strains, influenza A/WSN/33 and A/PR/8/34, which were two well characterized laboratory strains used in influenza studies. After validation of this method with laboratory strains, we can potentially apply this methodology to the surveillance of novel and more pathogenic strains, or drug resistant strains, such as H1N1 swine flu in year 2009, and the seasonal A/H1N1 flu.

The combination of base-specific cleavage and MALDI-MS provides a powerful platform for the large-scale comparative sequence analysis. At a laser repetition rate of 20 Hz, a sample spot can be measured in about 5 seconds if the sample is rastered up to five times at different positions. The target length of the samples analyzed was about 500bps, and four cleavage reactions were performed in order to identify a maximum of mutations. In this case, a single mass spectrometer would be able to read over 2 megabases in 24 hours. More importantly, this approach offers collateral security: identification of most mutations is based on multiple reactions. When a sequence change happens in the target region, up to 10 observations of mass peak changes could be found in the related four spectra. This advantage of multiple independent verifications is unique in this approach.

We processed comparative sequencing of HA and NA genes from influenza A /WSN/33 and A/PR/8/34 viral strains. The mass peak patterns obtained from the experimental spectra were proved as a fingerprint in diagnostic study of a certain viral strain. When primers were designed in the conserved regions of HA or NA genes from different viral strains, the PCR

products from related viral strains would be similar in length but slightly different in the sequence. The differences in the sequence would finally result in a certain number of “unique” peaks from the base-specific cleavage & MALDI-TOF MS assay. These unique peaks could be used as a fingerprint of the related viral strain.

The base-specific cleavage & MALDI-TOF MS approach not only can be used for typing of known sequence changes, which provided a powerful platform for surveillance of those drug resistance related mutations, but also shows potential for detection and identification of unknown sequence changes.

To cover the full length of any gene from influenza, at most four PCR reactions are needed. If a mutation occurs, it could be at least located in a small region and in some cases, the exact location of the mutation could be identified.

The base-specific cleavage & MALDI MS assay could achieve more than 90% coverage in a comparative sequencing analysis for a sequence with 300~800 bases in length. Compared with conventional sequencing approach, the base-specific cleavage & MALDI MS assay could achieve higher throughput with lower cost, and more importantly, base-specific cleavage could offer more accurate result: in most of cases, the detection of a sequence change relied on multiple “observations”. This is an important advantage for diagnostic applications.

The major limit factor of the base-specific cleavage & MALDI-TOF MS assay is the size of the cleavage product. In our experiment, those fragments larger than 6000 Da, which were about 20 bases in length, were difficult to get ionized with the current used matrix. We have found a compound matrix mixed from 3-HPA and ATT that could increase the signal to noise

ratio and the resolution of certain mass range. Further optimization of the matrix for nucleic acid would increase the sequence coverage of the comparative sequencing approach using base-specific cleavage & MALDI MS assay.

2',3'-cyclic phosphodiester intermediate could always been found in the cleaved product from RNase T1 but not RNase A. However, in our experiment, we did observe some peaks which were most probably derived from 2',3'-cyclic phosphodiester intermediate in the spectra of a RNase A cleaved product. The cleavage product with a cyclic phosphate would have a 18 Da mass shift from the theoretical result, while a A/G mutation would introduce a 16 Da mass shift. These two shift peaks were partially overlapped thus a high resolution MS was needed in order to differentiate them. The existence of 2',3'-cyclic phosphodiester intermediate in RNase A cleaved products needs to be further investigated in our future study.

From the present study, it can be concluded that base-specific cleavage combined with MALDI-MS can be utilized as a powerful platform for the comparative sequencing of influenza HA and NA genes, and screening of viral mutations in these influenza strains. After validation of this method with laboratory strains, we can potentially apply this methodology in the surveillance of novel and more pathogenic strains, or drug resistant strains, such as H1N1 swine flu in year 2009, and the seasonal A/H1N1 flu.

References

1. Rassoulzadegan, M., Grandjean, V., Gounon, P., Vincent, S., Gillot, I. and Cuzin, F. (2006) RNA-mediated non-mendelian inheritance of an epigenetic change in the mouse. *Nature*, **441**, 469-474.
2. Pennisi, E. (2007) Genomics - DNA study forces rethink of what it means to be a gene. *Science*, **316**, 1556-1557.
3. Pearson, H. (2006) What is a gene? *Nature*, **441**, 398-401.
4. Noble, D. (2008) Genes and causation. *Philos T R Soc A*, **366**, 3001-3015.
5. Brookes, A.J. (1999) The essence of SNPs. *Gene*, **234**, 177-186.
6. Waikan, Y. and Dozy, A.M. (1978) Polymorphism of DNA-Sequence Adjacent to Human Beta-Globin Structural Gene - Relationship to Sickle Mutation. *PNAS*, **75**, 5631-5635.
7. Syvanen, A.C. (2001) Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nat Rev Genet*, **2**, 930-942.
8. Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M. and Sirotkin, K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*, **29**, 308-311.
9. Thorisson, G.A., Smith, A.V., Krishnan, L. and Stein, L.D. (2005) The International HapMap Project Web site. *Genome Res*, **15**, 1592-1593.
10. Fredman, D., Siegfried, M., Yuan, Y.P., Bork, P., Lehvaslaiho, H. and Brookes, A.J. (2002) HGVbase: a human sequence variation database emphasizing data quality and a broad spectrum of data sources. *Nucleic Acids Res*, **30**, 387-391.
11. Hinds, D.A., Stuve, L.L., Nilsen, G.B., Halperin, E., Eskin, E., Ballinger, D.G., Frazer, K.A. and Cox, D.R. (2005) Whole-genome patterns of common DNA variation in three human populations. *Science*, **307**, 1072-1079.
12. Hirakawa, M., Tanaka, T., Hashimoto, Y., Kuroda, M., Takagi, T. and Nakamura, Y. (2002) JSNP: a database of common gene variations in the Japanese population. *Nucleic Acids Res*, **30**, 158-162.
13. Dooley, E.E. (2003) The Pharmacogenetics Research Network and the Pharmacogenetics and Pharmacogenomics Knowledge Base. *Environ Health Persp*, **111**, A575-A575.
14. Hernandez-Boussard, T., Whirl-Carrillo, M., Hebert, J.M., Gong, L., Owen, R., Gong, M., Gor, W., Liu, F., Truong, C., Whaley, R. *et al.* (2008) The pharmacogenetics and pharmacogenomics knowledge base: accentuating the knowledge. *Nucleic Acids Res*, **36**, D913-D918.
15. Hewett, M., Oliver, D.E., Rubin, D.L., Easton, K.L., Stuart, J.M., Altman, R.B. and Klein, T.E. (2002) PharmGKB: The Pharmacogenetics Knowledge Base. *Nucleic Acids Res*, **30**, 163-165.
16. Packer, B.R., Yeager, M., Staats, B., Welch, R., Crenshaw, A., Kiley, M., Eckert, A., Beerman, M., Miller, E., Bergen, A. *et al.* (2004) SNP500Cancer: a public resource for sequence validation and assay development for genetic variation in candidate genes. *Nucleic Acids Res*, **32**, D528-D532.
17. Livingston, R.J., von Niederhausern, A., Jegga, A.G., Crawford, D.C., Carlson, C.S., Rieder, M.J., Gowrisankar, S., Aronow, B.J., Weiss, R.B. and Nickerson, D.A. (2004) Pattern of sequence variation across 213 environmental response genes. *Genome Res*, **14**, 1821-1831.
18. Ingelman-Sundberg, M., Oscarson, M., Daly, A.K., Garte, S. and Nebert, D.W. (2001) Human

- cytochrome P-450 (CYP) genes: A web page for the nomenclature of alleles. *Cancer Epidemiol Biomarkers*, **10**, 1307-1308.
19. Hollegaard, M.V. and Bidwell, J.L. (2006) Cytokine gene polymorphism in human disease: on-line databases, Supplement 3. *Genes Immun*, **7**, 269-276.
 20. Haukim, N., Bidwell, J.L., Smith, A.J.P., Keen, L.J., Gallagher, G., Kimberly, R., Huizinga, T., McDermott, M.F., Oksenberg, J., McNicholl, J. *et al.* (2002) Cytokine gene polymorphism in human disease: on-line databases, Supplement 2. *Genes Immun*, **3**, 313-330.
 21. Bidwell, J., Keen, L., Gallagher, G., Kimberly, R., Huizinga, T., McDermott, M.F., Oksenberg, J., McNicholl, J., Pociot, F., Hardt, C. *et al.* (2001) Cytokine gene polymorphism in human disease: on-line databases, Supplement 1. *Genes Immun*, **2**, 61-70.
 22. Bidwell, J., Keen, L., Gallagher, G., Kimberly, R., Huizinga, T., McDermott, M.F., Oksenberg, J., McNicholl, J., Pociot, F., Hardt, C. *et al.* (1999) Cytokine gene polymorphism in human disease: on-line databases. *Genes Immun*, **1**, 3-19.
 23. Schuler, G.D., Boguski, M.S., Stewart, E.A., Stein, L.D., Gyapay, G., Rice, K., White, R.E., RodriguezTome, P., Aggarwal, A., Bajorek, E. *et al.* (1996) A gene map of the human genome. *Science*, **274**, 540-546.
 24. Collins, F.S., Lander, E.S., Rogers, J., Waterston, R.H. and Conso, I.H.G.S. (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931-945.
 25. Sachidanandam, R., Weissman, D., Schmidt, S.C., Kakol, J.M., Stein, L.D., Marth, G., Sherry, S., Mullikin, J.C., Mortimore, B.J., Willey, D.L. *et al.* (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, **409**, 928-933.
 26. Wang, D.G., Fan, J.B., Siao, C.J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J. *et al.* (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science*, **280**, 1077-1082.
 27. Schmid, K.J., Sorensen, T.R., Stracke, R., Torjek, O., Altmann, T., Mitchell-Olds, T. and Weisshaar, B. (2003) Large-scale identification and analysis of genome-wide single-nucleotide polymorphisms for mapping in *Arabidopsis thaliana*. *Genome Res*, **13**, 1250-1257.
 28. Li, W.H. and Sadler, L.A. (1991) Low Nucleotide Diversity in Man. *Genetics*, **129**, 513-523.
 29. Nickerson, D.A., Taylor, S.L., Weiss, K.M., Clark, A.G., Hutchinson, R.G., Stengard, J., Salomaa, V., Vartiainen, E., Boerwinkle, E. and Sing, C.F. (1998) DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. *Nat Genet*, **19**, 233-240.
 30. Holden, A.L. (2002) The SNP consortium: summary of a private consortium effort to develop an applied map of the human genome. *Biotechniques*, **Suppl**, 22-24, 26.
 31. Kim, S. and Misra, A. (2007) SNP genotyping: Technologies and biomedical applications. *Annu Rev Biomed Eng*, **9**, 289-320.
 32. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) DNA sequencing with chain-terminating inhibitors. *PNAS*, **74**, 5463-5467.
 33. Dianzani, I., Camaschella, C., Ponzzone, A. and Cotton, R.G. (1993) Dilemmas and progress in mutation detection. *Trends Genet*, **9**, 403-405.
 34. Cotton, R.G. (1997) Slowly but surely towards better scanning for mutations. *Trends Genet*, **13**, 43-46.

35. Oberacher, H. (2008) On the use of different mass spectrometric techniques for characterization of sequence variability in genomic DNA. *Anal Bioanal Chem*, **391**, 135-149.
36. Kristensen, V.N., Kelefiotis, D., Kristensen, T. and Borresen-Dale, A.L. (2001) High-throughput methods for detection of genetic variation. *Biotechniques*, **30**, 318-322, 324, 326 passim.
37. Sobrino, B. and Carracedo, A. (2005) SNP typing in forensic genetics: a review. *Methods Mol Biol*, **297**, 107-126.
38. Karas, M., Bachmann, D. and Hillenkamp, F. (1985) Influence of the Wavelength in High-Irradiance Ultraviolet-Laser Desorption Mass-Spectrometry of Organic-Molecules. *Anal Chem*, **57**, 2935-2939.
39. Karas, M., Bachmann, D., Bahr, U. and Hillenkamp, F. (1987) Matrix-Assisted Ultraviolet-Laser Desorption of Nonvolatile Compounds. *Int J Mass Spectrom*, **78**, 53-68.
40. Tanaka, K.W., H.; Ido, Y.; Akita, S.; Yoshida, Y.; Yoshida, T. (1988) Protein and Polymer Analyses up to m/z 100 000 by Laser Ionization Time-of flight Mass Spectrometry. *Rapid Commun Mass Spectrom*, **20**, 151-153.
41. Markides, K.G., A. Advanced information on the Nobel Prize in Chemistry 2002.
42. Karas, M. and Hillenkamp, F. (1988) Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Anal Chem*, **60**, 2299-2301.
43. Karas, M., Bahr, U. and Hillenkamp, F. (1989) Uv Laser Matrix Desorption Ionization Mass-Spectrometry of Proteins in the 100 000 Dalton Range. *Int J Mass Spectrom*, **92**, 231-242.
44. Beavis, R.C. and Chait, B.T. (1989) Matrix-assisted laser-desorption mass spectrometry using 355 nm radiation. *Rapid Commun Mass Spectrom*, **3**, 436-439.
45. Fenn, J.B., Mann, M., Meng, C.K., Wong, S.F. and Whitehouse, C.M. (1989) Electrospray ionization for mass spectrometry of large biomolecules. *Science*, **246**, 64-71.
46. Zenobi, R. and Knochenmuss, R. (1998) Ion formation in MALDI mass spectrometry. *Mass Spectrom Rev*, **17**, 337-366.
47. F.Hillenkamp. (1983) Laser-induced ion formation from organic solids. *Springer Ser. Chem. Phys*, **25**, 190-205.
48. Hillenkamp, F., Karas, M., Beavis, R.C. and Chait, B.T. (1991) Matrix-assisted laser desorption/ionization mass spectrometry of biopolymers. *Anal Chem*, **63**, 1193A-1203A.
49. Michael Karas, U.B., Ulrich Gießmann. (1991) Matrix-assisted laser desorption ionization mass spectrometry. *Mass Spectrom Rev*, **10**, 335-357.
50. Ehring, H., Karas, M. and Hillenkamp, F. (1992) Role of Photoionization and Photochemistry in Ionization Processes of Organic-Molecules and Relevance for Matrix-Assisted Laser Desorption Ionization Mass-Spectrometry. *Org Mass Spectrom*, **27**, 472-480.
51. Dreisewerd, K. (2003) The desorption process in MALDI. *Chem Rev*, **103**, 395-425.
52. Karas, M. and Kruger, R. (2003) Ion formation in MALDI: The cluster ionization mechanism. *Chem Rev*, **103**, 427-439.
53. Knochenmuss, R. and Zenobi, R. (2003) MALDI ionization: The role of in-plume processes. *Chem Rev*, **103**, 441-452.
54. Spengler, B., Pan, Y., Cotter, R.J. and Kan, L.S. (1990) Molecular weight determination of underivatized oligodeoxyribonucleotides by positive-ion matrix-assisted ultraviolet laser-desorption mass spectrometry. *Rapid Commun Mass Spectrom*, **4**, 99-102.

55. Bornsen, K.O., Schar, M. and Widmer, H.M. (1990) Matrix-Assisted Laser Desorption and Ionization Mass-Spectrometry and Its Applications in Chemistry. *Chimia*, **44**, 412-416.
56. Karas, M. and Bahr, U. (1990) Laser Desorption Ionization Mass-Spectrometry of Large Biomolecules. *Trac-Trend Anal Chem*, **9**, 321-325.
57. Tang, K., Allman, S.L. and Chen, C.H. (1992) Mass-Spectrometry of Laser-Desorbed Oligonucleotides. *Rapid Commun Mass Spectrom*, **6**, 365-368.
58. Nordhoff, E., Ingendoh, A., Cramer, R., Overberg, A., Stahl, B., Karas, M., Hillenkamp, F. and Crain, P.F. (1992) Matrix-Assisted Laser Desorption Ionization Mass-Spectrometry of Nucleic-Acids with Wavelengths in the Ultraviolet and Infrared. *Rapid Commun Mass Spectrom*, **6**, 771-776.
59. Wu, K.J., Steding, A. and Becker, C.H. (1993) Matrix-Assisted Laser Desorption Time-of-Flight Mass-Spectrometry of Oligonucleotides Using 3-Hydroxypicolinic Acid as an Ultraviolet-Sensitive Matrix. *Rapid Commun Mass Spectrom*, **7**, 142-146.
60. Pieleles, U., Zurcher, W., Schar, M. and Moser, H.E. (1993) Matrix-assisted laser desorption ionization time-of-flight mass spectrometry: a powerful tool for the mass and sequence analysis of natural and modified oligonucleotides. *Nucleic Acids Res*, **21**, 3191-3196.
61. Zhu, Y.F., Chung, C.N., Taranenko, N.I., Allman, S.L., Martin, S.A., Haff, L. and Chen, C.H. (1996) The study of 2,3,4-trihydroxyacetophenone and 2,4,6-trihydroxyacetophenone as matrices for DNA detection in matrix-assisted laser desorption ionization time-of-flight mass spectrometry. *Rapid Commun Mass Spectrom*, **10**, 383-388.
62. Tang, K., Taranenko, N.I., Allman, S.L., Chen, C.H., Chang, L.Y. and Jacobson, K.B. (1994) Picolinic-Acid as a Matrix for Laser Mass-Spectrometry of Nucleic-Acids and Proteins. *Rapid Commun Mass Spectrom*, **8**, 673-677.
63. Tang, K., Taranenko, N.I., Allman, S.L., Chang, L.Y. and Chen, C.H. (1994) Detection of 500-Nucleotide DNA by Laser-Desorption Mass-Spectrometry. *Rapid Commun Mass Spectrom*, **8**, 727-730.
64. Taranenko, N.I., Tang, K., Allman, S.L., Ch'ang, L.Y. and Chen, C.H. (1994) 3-Aminopicolinic acid as a matrix for laser desorption mass spectrometry of biopolymers. *Rapid Commun Mass Spectrom*, **8**, 1001-1006.
65. Lecchi, P. and Pannell, L.K. (1995) The Detection of Intact Double-Stranded DNA by Maldi. *J Am Soc Mass Spectr*, **6**, 972-975.
66. Distler, A.M. and Allison, J. (2001) 5-methoxysalicylic acid and spermine: A new matrix for the matrix-assisted laser desorption/ionization mass spectrometry analysis of oligonucleotides. *J Am Soc Mass Spectr*, **12**, 456-462.
67. Song, F.H. (2003) Quinaldic acid as a new matrix for matrix-assisted laser desorption/ionization of nucleic acids. *Rapid Commun Mass Spectrom*, **17**, 1802-1807.
68. Zhou, L.H., Deng, H.M., Deng, Q.Y. and Zhao, S.K. (2004) A mixed matrix of 3-hydroxypicolinic acid and pyrazinecarboxylic acid for matrix-assisted laser desorption/ionization time-of-flight mass spectrometry of oligodeoxynucleotides. *Rapid Commun Mass Spectrom*, **18**, 787-794.
69. Zhang, Z.Y., Zhou, L.H., Zhao, S.K., Deng, H.M. and Deng, Q.Y. (2006) 3-hydroxycoumarin as a new matrix for matrix-assisted laser desorption/ionization time-of-flight mass spectrometry of DNA. *J Am Soc Mass Spectr*, **17**, 1665-1668.
70. Yu, F., Xu, S.Y., Pan, C.S., Ye, M.L., Zou, H.F. and Guo, B.C. (2006) A matrix of

- 3,4-diaminobenzophenone for the analysis of oligonucleotides by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Nucleic Acids Res*, **34**, e94.
71. Berkenkamp, S., Kirpekar, F. and Hillenkamp, F. (1998) Infrared MALDI mass spectrometry of large nucleic acids. *Science*, **281**, 260-262.
 72. Currie, G.J. and Yates, J.R. (1993) Analysis of Oligodeoxynucleotides by Negative-Ion Matrix-Assisted Laser-Desorption Mass-Spectrometry. *J Am Soc Mass Spectr*, **4**, 955-963.
 73. Li, Y.C.L., Cheng, S.W. and Chan, T.W.D. (1998) Evaluation of ammonium salts as co-matrices for matrix-assisted laser desorption/ionization mass spectrometry of oligonucleotides. *Rapid Commun Mass Spectrom*, **12**, 993-998.
 74. Asara, J.M. and Allison, J. (1999) Enhanced detection of oligonucleotides in UV MALDI MS using the tetraamine spermine as a matrix additive. *Anal Chem*, **71**, 2866-2870.
 75. Shahgholi, M., Garcia, B.A., Chiu, N.H.L., Heaney, P.J. and Tang, K. (2001) Sugar additives for MALDI matrices improve signal allowing the smallest nucleotide change (A : T) in a DNA sequence to be resolved. *Nucleic Acids Res*, **29**, e91.
 76. Garcia, B.A., Heaney, P.J. and Tang, K. (2002) Improvement of the MALDI-TOF analysis of DNA with thin-layer matrix preparation. *Anal Chem*, **74**, 2083-2091.
 77. Tang, K., Opalsky, D., Abel, K., van den Boom, D., Yip, P., Del Mistro, G., Braun, A. and Cantor, C.R. (2003) Single nucleotide polymorphism analyses by MALDI-TOF MS. *Int J Mass Spectrom*, **226**, 37-54.
 78. Kirpekar, F., Nordhoff, E., Larsen, L.K., Kristiansen, K., Roepstorff, P. and Hillenkamp, F. (1998) DNA sequence analysis by MALDI mass spectrometry. *Nucleic Acids Res*, **26**, 2554-2559.
 79. Gilar, M., Belenky, A. and Wang, B.H. (2001) High-throughput biopolymer desalting by solid-phase extraction prior to mass spectrometric analysis. *J Chromatogr A*, **921**, 3-13.
 80. Nordhoff, E., Kirpekar, F. and Roepstorff, P. (1996) Mass spectrometry of nucleic acids. *Mass Spectrom Rev*, **15**, 67-138.
 81. Tang, K., Shahgholi, M., Garcia, B.A., Heaney, P.J., Cantor, C.R., Scott, L.G. and Williamson, J.R. (2002) Improvement in the apparent mass resolution of oligonucleotides by using C-12/N-14-enriched samples. *Anal Chem*, **74**, 226-231.
 82. Li, Y.Z., Tang, K., Little, D.P., Koster, H., Hunter, R.L. and McIver, R.T. (1996) High-resolution MALDI Fourier transform mass spectrometry of oligonucleotides. *Anal Chem*, **68**, 2090-2096.
 83. Braun, A., Little, D.P. and Koster, H. (1997) Detecting CFTR gene mutations by using primer oligo base extension and mass spectrometry. *Clin Chem*, **43**, 1151-1158.
 84. Haff, L.A. and Smirnov, I.P. (1997) Single-nucleotide polymorphism identification assays using a thermostable DNA polymerase and delayed extraction MALDI-TOF mass spectrometry. *Genome Res*, **7**, 378-388.
 85. Ross, P., Hall, L., Smirnov, I. and Haff, L. (1998) High level multiplex genotyping by MALDI-TOF mass spectrometry. *Nat Biotechnol*, **16**, 1347-1351.
 86. Sequenom. (2009), *iPLEX® Gold Application Guide*, San Diego.
 87. Wang, L., Luhm, R. and Lei, M. (2007) SNP and mutation analysis. *Adv Exp Med Biol*, **593**, 105-116.
 88. Vogel, N., Schiebel, K. and Humeny, A. (2009) Technologies in the Whole-Genome Age: MALDI-TOF-Based Genotyping. *Transfus Med Hemoth*, **36**, 253-262.

89. Jaremko, M., Justenhoven, C., Abraham, B.K., Schroth, W., Fritz, P., Brod, S., Vollmert, C., Illig, T. and Brauch, H. (2005) MALDI-TOF MS and TaqMan((R)) assisted SNP genotyping of DNA isolated from formalin-fixed and paraffin-embedded tissues (FFPET). *Hum Mutat*, **25**, 232-238.
90. Horn, H., Pott, C., Kalla, J., Dreyling, M., Rosenwald, A., Ott, G., Schwab, M. and Schaeffeler, E. (2010) A multiplex MALDI-TOF MS approach facilitates genotyping of DNA from formalin-fixed paraffin-embedded tumour specimens. *Pharmacogenet Genom*, **20**, 598-604.
91. Soderlund-Strand, A., Dillner, J. and Carlson, J. (2008) High-throughput genotyping of oncogenic human papilloma viruses with MALDI-TOF mass spectrometry. *Clin Chem*, **54**, 86-92.
92. Cerezo, M., Cerny, V., Carracedo, A. and Salas, A. (2009) Applications of MALDI-TOF MS to large-scale human mtDNA population-based studies. *Electrophoresis*, **30**, 3665-3673.
93. Thompson, S.L., Lamothe, M., Meirmans, P.G., Perinet, P. and Isabel, N. (2010) Repeated unidirectional introgression towards *Populus balsamifera* in contact zones of exotic and native poplars. *Mol Ecol*, **19**, 132-145.
94. Thomas, R.K., Baker, A.C., DeBiasi, R.M., Winckler, W., LaFramboise, T., Lin, W.M., Wang, M., Feng, W., Zander, T., MacConnaill, L.E. *et al.* (2007) High-throughput oncogene mutation profiling in human cancer (vol 39, pg 347, 2007). *Nat Genet*, **39**, 567-567.
95. Lubezky, N., Ben-Haim, M., Marmor, S., Brazowsky, E., Rechavi, G., Klausner, J.M. and Cohen, Y. (2011) High-Throughput Mutation Profiling in Intraductal Papillary Mucinous Neoplasm (IPMN). *J Gastrointest Surg*, **15**, 503-511.
96. Kiessling, M.K., Oberholzer, P.A., Mondal, C., Karpova, M.B., Zipser, M.C., Lin, W.M., Girardi, M., MacConaill, L.E., Kehoe, S.M., Hatton, C. *et al.* (2011) High-throughput mutation profiling of CTCL samples reveals KRAS and NRAS mutations sensitizing tumors toward inhibition of the RAS/RAF/MEK signaling cascade. *Blood*, **117**, 2433-2440.
97. Huebner, C., Petermann, I., Browning, B.L., Shelling, A.N. and Ferguson, L.R. (2007) Triallelic single nucleotide polymorphisms and genotyping error in genetic epidemiology studies: MDR1 (ABCB1) G2677/T/A as an example. *Cancer Epidem Biomar*, **16**, 1185-1192.
98. Sasayama, T., Kato, M., Aburatani, H., Kuzuya, A. and Komiyama, M. (2006) Simultaneous genotyping of indels and SNPs by mass spectroscopy. *J Am Soc Mass Spectr*, **17**, 3-8.
99. Risch, N.J. (2000) Searching for genetic determinants in the new millennium. *Nature*, **405**, 847-856.
100. Goldstein, D.B., Ahmadi, K.R., Weale, M.E. and Wood, N.W. (2003) Genome scans and candidate gene approaches in the study of common diseases and variable drug responses. *Trends Genet*, **19**, 615-622.
101. Kammerer, S., Roth, R.B., Reneland, R., Marnellos, G., Hoyal, C.R., Markward, N.J., Ebner, F., Kiechle, M., Schwarz-Boeger, U., Griffiths, L.R. *et al.* (2004) Large-scale association study identifies ICAM gene region as breast and prostate cancer susceptibility locus. *Cancer Res*, **64**, 8906-8910.
102. Suttner, K., Rosenstiel, P., Depner, M., Schedel, M., Pinto, L.A., Ruether, A., Adamski, J., Klopp, N., Illig, T., Vogelberg, C. *et al.* (2009) TBX21 gene variants increase childhood asthma risk in combination with HLX1 variants. *J Allergy Clin Immun*, **123**, 1062-1068.
103. Abel, K., Kammerer, S., Hoyal, C., Reneland, R., Marnellos, G., Nelson, M.R. and Braun, A. (2005) Identification of susceptibility genes and genetic modifiers of human diseases. *P Soc*

- Photo-Opt Ins*, **5699**, 491-502,542.
104. Abel, K., Reneland, R., Kammerer, S., Mah, S., Hoyal, C., Cantor, C.R., Nelson, M.R. and Braun, A. (2006) Genome-wide SNP association: Identification of susceptibility alleles for osteoarthritis. *Autoimmun Rev*, **5**, 258-263.
 105. Volkman, S.K., Sabeti, P.C., DeCaprio, D., Neafsey, D.E., Schaffner, S.F., Milner, D.A., Daily, J.P., Sarr, O., Ndiaye, D., Ndir, O. *et al.* (2007) A genome-wide map of diversity in *Plasmodium falciparum*. *Nat Genet*, **39**, 113-119.
 106. Herring, C.D., Raghunathan, A., Honisch, C., Patel, T., Applebee, M.K., Joyce, A.R., Albert, T.J., Blattner, F.R., van den Boom, D., Cantor, C.R. *et al.* (2006) Comparative genome sequencing of *Escherichia coli* allows observation of bacterial evolution on a laboratory timescale. *Nat Genet*, **38**, 1406-1412.
 107. Ding, C.M. and Cantor, C.R. (2003) A high-throughput gene expression analysis technique using competitive PCR and matrix-assisted laser desorption ionization time-of-flight MS. *PNAS*, **100**, 3059-3064.
 108. Rode, T.M., Berget, I., Langsrud, S., Moretro, T. and Holck, A. (2009) MALDI-TOF mass spectrometry for quantitative gene expression analysis of acid responses in *Staphylococcus aureus*. *J Microbiol Meth*, **78**, 86-93.
 109. Duffield, D.S., Cai, L. and Kim, S. (2010) Simultaneous determination of multiple mRNA levels utilizing MALDI-TOF mass spectrometry and biotinylated dideoxynucleotides. *Rna*, **16**, 1285-1291.
 110. Ding, C.M., Chiu, R.W.K., Lau, T.K., Leung, T.N., Chan, L.C., Chan, A.Y.Y., Charoenkwan, P., Ng, I.S.L., Law, H.Y., Ma, E.S.K. *et al.* (2004) MS analysis of single-nucleotide differences in circulating nucleic acids: Application to noninvasive prenatal diagnosis. *PNAS*, **101**, 10762-10767.
 111. Daelemans, C., Ritchie, M.E., Smits, G., Abu-Amero, S., Sudbery, I.M., Forrest, M.S., Campino, S., Clark, T.G., Stanier, P., Kwiatkowski, D. *et al.* (2010) High-throughput analysis of candidate imprinted genes and allele-specific gene expression in the human term placenta. *Bmc Genet*, **11**, 25.
 112. Lo, Y.M.D., Tsui, N.B.Y., Chiu, R.W.K., Lau, T.K., Leung, T.N., Heung, M.M.S., Gerovassili, A., Jin, Y.J., Nicolaides, K.H., Cantor, C.R. *et al.* (2007) Plasma placental RNA allelic ratio permits noninvasive prenatal chromosomal aneuploidy detection. *Nat Med*, **13**, 218-223.
 113. Zhong, X.Y. and Holzgreve, W. (2009) MALDI-TOF MS in Prenatal Genomics. *Transfus Med Hemoth*, **36**, 263-272.
 114. Mathias Ehrich, F.H., Dirk van den boom. (2007), *Perspectives in Bioanalysis*. Elsevier, Vol. 2, pp. 97-117.
 115. Fan, A.X.C., Garritsen, H.S.P., Tarhouny, S.E.L., Morris, M., Hahn, S., Holzgreve, W. and Zhong, X.Y. (2008) A rapid and accurate approach to identify single nucleotide polymorphisms of mitochondrial DNA using MALDI-TOF mass spectrometry. *Clin Chem Lab Med*, **46**, 299-305.
 116. Bocker, S. (2003) SNP and mutation discovery using base-specific cleavage and MALDI-TOF mass spectrometry. *Bioinformatics*, **19**, i44-i53.
 117. Tang, K., Oeth, P., Kammerer, S., Denissenko, M.F., Ekblom, J., Jurinke, C., van den Boom, D., Braun, A. and Cantor, C.R. (2004) Mining disease susceptibility genes through SNP analyses and expression profiling using MALDI-TOF mass spectrometry. *J Proteome Res*, **3**,

- 218-227.
118. Krebs, S., Medugorac, I., Seichter, D. and Forster, M. (2003) RNaseCut: a MALDI mass spectrometry-based method for SNP discovery. *Nucleic Acids Res*, **31**, e37.
 119. Stanssens, P., Zabeau, M., Meersseman, G., Remes, G., Gansemans, Y., Storm, N., Hartmer, R., Honisch, C., Rodi, C.P., Bocker, S. *et al.* (2004) High-throughput MALDI-TOF discovery of genomic sequence polymorphisms. *Genome Res*, **14**, 126-133.
 120. Henzel, W.J., Billeci, T.M., Stults, J.T., Wong, S.C., Grimley, C. and Watanabe, C. (1993) Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. *PNAS*, **90**, 5011-5015.
 121. Mann, M., Hojrup, P. and Roepstorff, P. (1993) Use of Mass-Spectrometric Molecular-Weight Information to Identify Proteins in Sequence Databases. *Biol Mass Spectrom*, **22**, 338-345.
 122. Pappin, D.J.C., Hojrup, P. and Bleasby, A.J. (1993) Rapid Identification of Proteins by Peptide-Mass Fingerprinting. *Curr Biol*, **3**, 327-332.
 123. Matthiesen, R. and Kirpekar, F. (2009) Identification of RNA molecules by specific enzyme digestion and mass spectrometry: software for and implementation of RNA mass mapping. *Nucleic Acids Res*, **37**, e48.
 124. Bocker, S. (2007) Simulating multiplexed SNP discovery rates using base-specific cleavage and mass spectrometry. *Bioinformatics*, **23**, E5-E12.
 125. Hossain, M. and Limbach, P.A. (2007) Mass spectrometry-based detection of transfer RNAs by their signature endonuclease digestion products. *Rna*, **13**, 295-303.
 126. Benitez-Paez, A., Villarroya, M., Douthwaite, S., Gabaldon, T. and Armengod, M.E. (2010) YibK is the 2'-O-methyltransferase TrmL that modifies the wobble nucleotide in Escherichia coli tRNA(Leu) isoacceptors. *Rna*, **16**, 2131-2143.
 127. Purta, E., Kaminska, K.H., Kasprzak, J.M., Bujnicki, J.M. and Douthwaite, S. (2008) YbeA is the m(3)Psi methyltransferase RlmH that targets nucleotide 1915 in 23S rRNA. *Rna*, **14**, 2234-2244.
 128. Purta, E., O'Connor, M., Bujnicki, J.M. and Douthwaite, S. (2009) YgdE is the 2'-O-ribose methyltransferase RlmM specific for nucleotide C2498 in bacterial 23S rRNA. *Mol Microbiol*, **72**, 1147-1158.
 129. Douthwaite, S. and Kirpekar, F. (2007) Identifying modifications in RNA by maldi mass spectrometry. *Method Enzymol*, **425**, 3-20.
 130. Kirpekar, F., Douthwaite, S. and Roepstorff, P. (2000) Mapping posttranscriptional modifications in 5S ribosomal RNA by MALDI mass spectrometry. *Rna*, **6**, 296-306.
 131. Mengel-Jorgensen, J. and Kirpekar, F. (2002) Detection of pseudouridine and other modifications in tRNA by cyanoethylation and MALDI mass spectrometry. *Nucleic Acids Res*, **30**, e135.
 132. Durairaj, A. and Limbach, P.A. (2008) Improving CMC-derivatization of pseudouridine in RNA for mass spectrometric detection. *Anal Chim Acta*, **612**, 173-181.
 133. Auxilien, S., Rasmussen, A., Rose, S., Brochier-Armanet, C., Husson, C., Fourmy, D., Grosjean, H. and Douthwaite, S. (2011) Specificity shifts in the rRNA and tRNA nucleotide targets of archaeal and bacterial m(5)U methyltransferases. *Rna*, **17**, 45-53.
 134. Goll, M.G., Kirpekar, F., Maggert, K.A., Yoder, J.A., Hsieh, C.L., Zhang, X.Y., Golic, K.G., Jacobsen, S.E. and Bestor, T.H. (2006) Methylation of tRNA(Asp) by the DNA methyltransferase homolog Dnmt2. *Science*, **311**, 395-398.

135. Schatz, P., Dietrich, D. and Schuster, M. (2004) Rapid analysis of CpG methylation patterns using RNase T1 cleavage and MALDI-TOF. *Nucleic Acids Res*, **32**, e167.
136. Dodge, J.E., Ramsahoye, B.H., Wo, Z.G., Okano, M. and Li, E. (2002) De novo methylation of MMLV provirus in embryonic stem cells: CpG versus non-CpG methylation. *Gene*, **289**, 41-48.
137. Haines, T.R., Rodenhiser, D.I. and Ainsworth, P.J. (2001) Allele-specific non-CpG methylation of the Nf1 gene during early mouse development. *Dev Biol*, **240**, 585-598.
138. Lister, R., Pelizzola, M., Downen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.M. *et al.* (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, **462**, 315-322.
139. Radpour, R., Haghighi, M.M., Fan, A.X.C., Torbati, P.M., Hahn, S., Holzgreve, W. and Zhong, X.Y. (2008) High-Throughput Hacking of the Methylation Patterns in Breast Cancer by In vitro Transcription and Thymidine-Specific Cleavage Mass Array on MALDI-TOF Silico-Chip. *Mol Cancer Res*, **6**, 1702-1709.
140. Ehrich, M., Nelson, M.R., Stanssens, P., Zabeau, M., Liloglou, T., Xinarianos, G., Cantor, C.R., Field, J.K. and van den Boom, D. (2005) Quantitative high-throughput analysis of DNA methylation patterns by base-specific cleavage and mass spectrometry. *PNAS*, **102**, 15785-15790.
141. Ehrich, M., Turner, J., Gibbs, P., Lipton, L., Giovanneti, M., Cantor, C. and van den Boom, D. (2008) Cytosine methylation profiling of cancer cell lines. *PNAS*, **105**, 4844-4849.
142. Coolen, M.W., Statham, A.L., Gardiner-Garden, M. and Clark, S.J. (2007) Genomic profiling of CpG methylation and allelic specificity using quantitative high-throughput mass spectrometry: critical evaluation and improvements. *Nucleic Acids Res*, **35**, e119.
143. Maiden, M.C.J., Bygraves, J.A., Feil, E., Morelli, G., Russell, J.E., Urwin, R., Zhang, Q., Zhou, J.J., Zurth, K., Caugant, D.A. *et al.* (1998) Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. *PNAS*, **95**, 3140-3145.
144. Urwin, R. and Maiden, M.C.J. (2003) Multi-locus sequence typing: a tool for global epidemiology. *Trends Microbiol*, **11**, 479-487.
145. Maiden, M.C.J. (2006) Multilocus sequence typing of bacteria. *Annu Rev Microbiol*, **60**, 561-588.
146. Levin, B.R., Lipsitch, M. and Bonhoeffer, S. (1999) Evolution and disease - Population biology, evolution, and infectious disease: Convergence and synthesis. *Science*, **283**, 806-809.
147. Musser, J.M. (1996) Molecular population genetic analysis of emerged bacterial pathogens: Selected insights. *Emerg Infect Dis*, **2**, 1-17.
148. Spratt, B.G. and Maiden, M.C.J. (1999) Bacterial population genetics, evolution and epidemiology. *Philos T R Soc A*, **354**, 701-710.
149. Enright, M.C. and Spratt, B.G. (1999) Multilocus sequence typing. *Trends Microbiol*, **7**, 482-487.
150. Maiden, M.C.J. (2000) High-throughput sequencing in the population analysis of bacterial pathogens of humans. *Int J Med Microbiol*, **290**, 183-190.
151. Jefferies, J., Clarke, S.C., Diggle, M.A., Smith, A., Dowson, C. and Mitchell, T. (2003) Automated pneumococcal MLST using liquid-handling robotics and a capillary DNA sequencer. *Mol Biotechnol*, **24**, 303-308.

152. Underwood, A. and Green, J. (2011) Call for a quality standard for sequence-based assays in clinical microbiology: necessity for quality assessment of sequences used in microbial identification and typing. *J Clin Microbiol*, **49**, 23-26.
153. Honisch, C., Chen, Y., Mortimer, C., Arnold, C., Schmidt, O., van den Boom, D., Cantor, C.R., Shah, H.N. and Gharbia, S.E. (2007) Automated comparative sequence analysis by base-specific cleavage and mass spectrometry for nucleic acid-based microbial typing. *PNAS*, **104**, 10649-10654.
154. Amann, R., Fuchs, B.M. and Behrens, S. (2001) The identification of microorganisms by fluorescence in situ hybridisation. *Curr Opin Biotech*, **12**, 231-236.
155. Clarridge, J.E. (2004) Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clin Microbiol Rev*, **17**, 840-862.
156. Jackson, G.W., McNichols, R.J., Fox, G.E. and Willson, R.C. (2007) Universal bacterial identification by mass spectrometry of 16S ribosomal RNA cleavage products. *Int J Mass Spectrom*, **261**, 218-226.
157. von Wintzingerode, F., Bocker, S., Schlotelburg, C., Chiu, N.H.L., Storm, N., Jurinke, C., Cantor, C.R., Gobel, U.B. and van den Boom, D. (2002) Base-specific fragmentation of amplified 16S rRNA genes analyzed by mass spectrometry: A tool for rapid bacterial identification. *PNAS*, **99**, 7039-7044.
158. Zhang, Z.D., Jackson, G.W., Fox, G.E. and Willson, R.C. (2006) Microbial identification by mass cataloging. *Bmc Bioinformatics*, **7**, 117.
159. Hartmer, R., Storm, N., Boecker, S., Rodi, C.P., Hillenkamp, F., Jurinke, C. and van den Boom, D. (2003) RNase T1 mediated base-specific cleavage and MALDI-TOF MS for high-throughput comparative sequence analysis. *Nucleic Acids Res*, **31**, e47.
160. Lefmann, M., Honisch, C., Bocker, S., Storm, N., von Wintzingerode, F., Schlotelburg, C., Moter, A., van den Boom, D. and Gobel, U.B. (2004) Novel mass spectrometry-based tool for genotypic identification of mycobacteria. *J Clin Microbiol*, **42**, 339-346.
161. Ganova-Raeva, L., Ramachandran, S., Honisch, C., Forbi, J.C., Zhai, X. and Khudyakov, Y. (2010) Robust Hepatitis B Virus Genotyping by Mass Spectrometry. *J Clin Microbiol*, **48**, 4161-4168.
162. Ruest, A., Michaud, S., Deslandes, S. and Frost, E.H. (2003) Comparison of the directigen Flu A+B test, the QuickVue influenza test, and clinical case definition to viral culture and reverse transcription-PCR for rapid diagnosis of influenza virus infection. *J Clin Microbiol*, **41**, 3487-3493.
163. van Elden, L.J., Nijhuis, M., Schipper, P., Schuurman, R. and van Loon, A.M. (2001) Simultaneous detection of influenza viruses A and B using real-time quantitative PCR. *J Clin Microbiol*, **39**, 196-200.
164. Wright, K.E., Wilson, G.A., Novosad, D., Dimock, C., Tan, D. and Weber, J.M. (1995) Typing and subtyping of influenza viruses in clinical samples by PCR. *J Clin Microbiol*, **33**, 1180-1184.
165. Schweiger, B., Zadow, I., Heckler, R., Timm, H. and Pauli, G. (2000) Application of a fluorogenic PCR assay for typing and subtyping of influenza viruses in respiratory samples. *J Clin Microbiol*, **38**, 1552-1558.
166. Barr, I.G., Komadina, N., Hurt, A.C., Iannello, P., Tomasov, C., Shaw, R., Durrant, C., Sjogren, H. and Hampson, A.W. (2005) An influenza A(H3) reassortant was epidemic in Australia and

- New Zealand in 2003. *J Med Virol*, **76**, 391-397.
167. Dharan, N.J., Gubareva, L.V., Meyer, J.J., Okomo-Adhiambo, M., McClinton, R.C., Marshall, S.A., George, K.S., Epperson, S., Brammer, L., Klimov, A.I. *et al.* (2009) Infections With Oseltamivir-Resistant Influenza A(H1N1) Virus in the United States. *JAMA*, **301**, 1034-1041.
 168. (1998) Randomised trial of efficacy and safety of inhaled zanamivir in treatment of influenza A and B virus infections. The MIST (Management of Influenza in the Southern Hemisphere Trialists) Study Group. *Lancet*, **352**, 1877-1881.
 169. Treanor, J.J., Hayden, F.G., Vrooman, P.S., Barbarash, R., Bettis, R., Riff, D., Singh, S., Kinnersley, N., Ward, P. and Mills, R.G. (2000) Efficacy and safety of the oral neuraminidase inhibitor oseltamivir in treating acute influenza: a randomized controlled trial. US Oral Neuraminidase Study Group. *JAMA*, **283**, 1016-1024.
 170. Baden, L.R., Drazen, J.M., Kritek, P.A., Curfman, G.D., Morrissey, S. and Campion, E.W. (2009) H1N1 influenza A disease--information for health professionals. *N Engl J Med*, **360**, 2666-2667.
 171. Dawood, F.S., Jain, S., Finelli, L., Shaw, M.W., Lindstrom, S., Garten, R.J., Gubareva, L.V., Xu, X., Bridges, C.B. and Uyeki, T.M. (2009) Emergence of a novel swine-origin influenza A (H1N1) virus in humans. *N Engl J Med*, **360**, 2605-2615.
 172. Smith, G.J., Vijaykrishna, D., Bahl, J., Lycett, S.J., Worobey, M., Pybus, O.G., Ma, S.K., Cheung, C.L., Raghwani, J., Bhatt, S. *et al.* (2009) Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature*, **459**, 1122-1125.
 173. Subbarao, K. and Joseph, T. (2007) Scientific barriers to developing vaccines against avian influenza viruses. *Nat Rev Immunol*, **7**, 267-278.
 174. Donald Voet, J.G.V. (2004) *Biochemistry, Third Edition*
 175. Ecker, D.J., Sampath, R., Massire, C., Blyn, L.B., Hall, T.A., Eshoo, M.W. and Hofstadler, S.A. (2008) Ibis T5000: a universal biosensor approach for microbiology. *Nat Rev Microbiol*, **6**, 553-558.
 176. Soundararajan, V., Tharakaraman, K., Raman, R., Raguram, S., Shriver, Z., Sasisekharan, V. and Sasisekharan, R. (2009) Extrapolating from sequence--the 2009 H1N1 'swine' influenza virus. *Nat Biotechnol*, **27**, 510-513.
 177. Deyde, V.M., Xu, X., Bright, R.A., Shaw, M., Smith, C.B., Zhang, Y., Shu, Y., Gubareva, L.V., Cox, N.J. and Klimov, A.I. (2007) Surveillance of resistance to adamantanes among influenza A(H3N2) and A(H1N1) viruses isolated worldwide. *J Infect Dis*, **196**, 249-257.
 178. Dharan, N.J., Gubareva, L.V., Meyer, J.J., Okomo-Adhiambo, M., McClinton, R.C., Marshall, S.A., St George, K., Epperson, S., Brammer, L., Klimov, A.I. *et al.* (2009) Infections with oseltamivir-resistant influenza A(H1N1) virus in the United States. *JAMA*, **301**, 1034-1041.
 179. Nogues, M.V., Vilanova, M. and Cuchillo, C.M. (1995) Bovine pancreatic ribonuclease A as a model of an enzyme with multiple substrate binding sites. *Biochim Biophys Acta*, **1253**, 16-24.