Supporting Information

High-throughput and automatic typing via human papillomavirus identification map for cervical cancer screening and prognosis

Linglu Yi, [a,b] Xueqin Xu,[*a] Xuexia Lin,[b] Haifang Li, [b] Yuan Ma [a] and Jin-Ming Lin [*b]

[a] College of Chemistry and Chemical Engineering, Fuzhou University, Fuzhou 350108, China. Email: xxq@fzu.edu.cn

[b] Department of Chemistry, Beijing Key Laboratory of Microanalytical Methods and Instrumentation, The Key Laboratory of Bioorganic Phosphorus Chemistry & Chemical Biology, Tsinghua University, Beijing 100084, China. Fax/Tel: +86 10 62792343; Email: jmlin@mail.tsinghua.edu.cn

Contents

1. The design and evaluation of PCR-RFLP-MCE method

1.1 List of standard RFLP patterns for 47 kinds of HPV types

Typing decision was made by comparing our measured results of RFLP patterns with standard RFLP patterns. 47 kinds of HPV types were finally chosen as candidates (including all high-risk ones, low-risk ones and some undefined ones).

TABLES

Table S1. 47 Kinds of HPV Typing Based on PCR with PGMY09/11 Combined with RFLP Analysis with PstI, HaeIII, DdeI, and RsaI Enzymes

	Rsa I patterns	Hae III patterns	Dde I patterns	Pst I patterns
A	18(135,125,85,72)	18(455)		18(242,213)
	54(138,125,117,72)	54(217,127,108)		54(452)
B	32(216,161,72)	32(317,124,8)		
	44(221,161,72)	44(223,124,108)		
C	26(365,72,18)	26(455)		
	62(359,72,18)	62(232,217)		62(342,108)
	69(365,72,18)	69(223,183,49)		69(455)
	40(365,90)	40(447,8)		
	72(365,72,18)	72(220,211,24)		
D	73(201,161,96)	73(458)		
	34(186,161,96,15)	34(334,124)		
E	16(310,72,70)	16(444,8)		16(216,210,26)
	43(332,72,45)	43(331,118)		43(273,176)
	84(310,142)	84(346, 106)		84(452)
	90(310,139)	90(232,209,8)		90(449)
	82(310,73,72)	82(447,8)		82(455)
	45(338,72,48)	45(447,8)		45(242, 213)
	56(310,72,49)	56(275,166,8)		56(242, 207)

	67(310,72)			67(423, 26)
F	52(449)	52(258,183,8)	52(357,92)	52(423,26)
	91(455)	91(455)	91(357,98)	91(455)
	53(449)	53(232,217)	53(206,158,85)	53(449)
	30(449)	30(232,217)	30(291,158)	30(242,207)
	66(449)	66(449)	66(291,158)	66(207,150,66,26)
	59(455)	59(396,56)	59(455)	59(426,26)
	81(452)	81(127,121,108,96)		81(341,111)
G	33(236,102,72,39)		33(320,77,52)	
	42(242,135,72,26)		42(341,108)	
	11(216,135,72,26)	11(217,124,108)	11(447,2)	
	70(231,123,72,29)	70(232,117,106)	70(455)	
H	31(380,72)	31(328,124)	31(283,167,2)	
	85(383,72)	85(455)	85(297,158)	
	89(380,72)	89(325,127)	89(246,152,54)	
	51(380,72)	51(379,73)	51(362,90)	
	87(380,72)	87(232,209,8,3)	87(362,90)	
	83(380,72)		83(452)	
	71(380,72)	71(217,127,108)	71(320,132)	
	86(380,72)	86(343,106,3)	86(310,125,17)	
I	35(171,161,72,42)		35(294,135,23)	
	64*(181,161,72)		64*(211,151,87,9)	
	55*(165,161,72,57)		55*(112,111,101,85,46)	
	61(185,180,72,18)		61(455)	
J	39(260,123,72)		39(324,131)	39(330,125)
	68(260,85,72,38)		68(455)	68(455)
K	6(161,149,72,67)	6(217,124,108)	6(382,67)	6(449)
	13(175,135,73)		13(326, 67, 62)	13(242,213)
L	58(306,111,32)			

HPV 55* and HPV 64* are the subtype of HPV 44 and HPV 34, respectively. HPV type is indicated by Arabic numerals with their corresponding RFLP fragment sizes bracketed after.

1.2 Performance of PCR-RFLP-MCE method

PCR-RFLP-MCE typing method had good performance in Hela cells and Caski cells. Figure S2 showed the stability and reproducibility of our method.

1.2.1 Stability and reproducibility of PCR-RFLP-MCE method

FIGURES

Figure S1

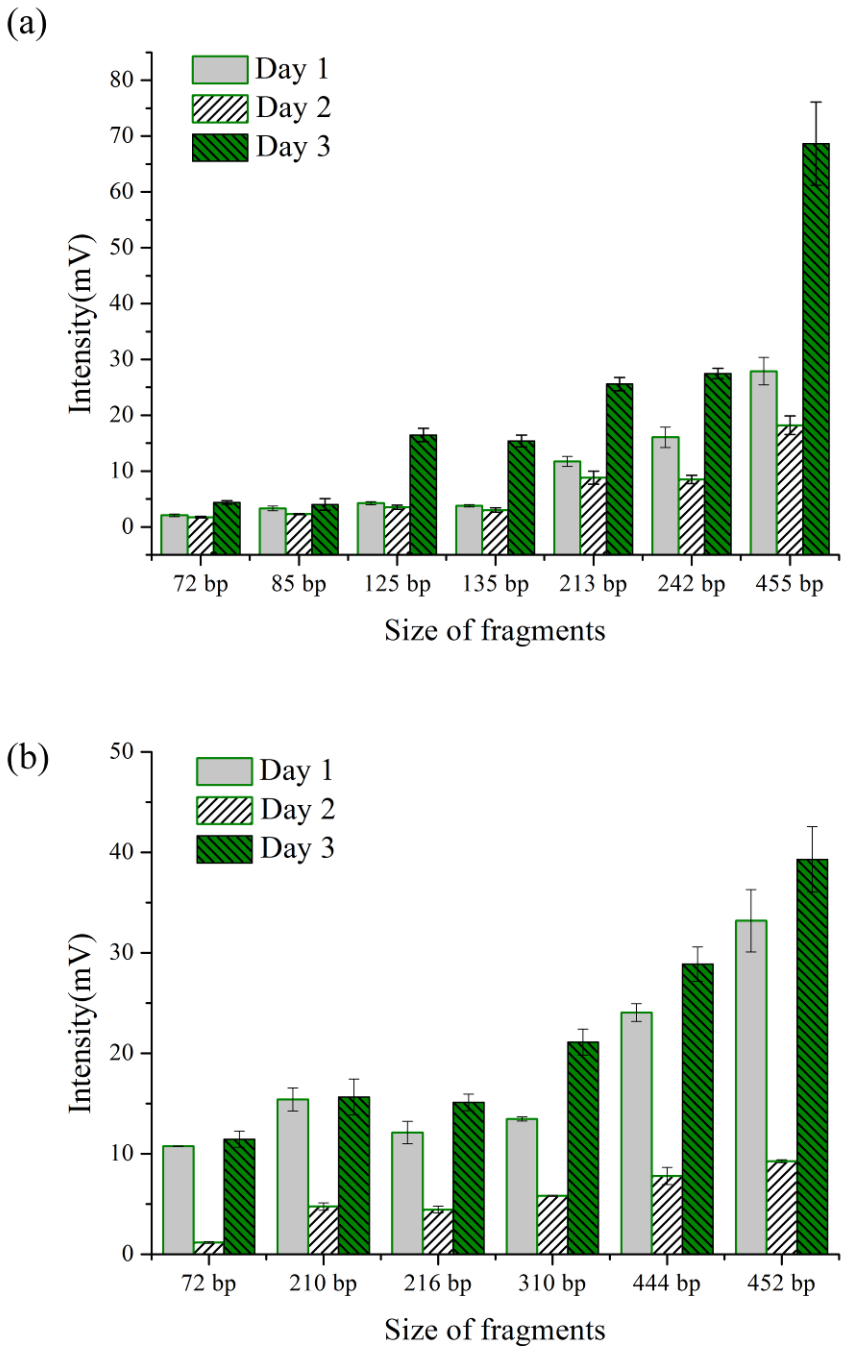

Figure S1. Repeatability was identified with Hela cells and Caski cells. a) RFLP fragments with 72, 85, 125, 135, 213, and 242 base pairs, and amplified template of 455 base pairs for Hela cells, (b) RFLP fragments of 72, 210, 216, 310 and 444 base pairs, and amplified template of 452 base pairs for Caski cells. Standard deviations for 3 replicates for each fragment in each day are shown as error bars.

1.2.2 Evaluation of results by calculation of Euclidean distance coefficient

Euclidean distant coefficient is a parameter used to evaluate the variation degree among different batches of samples.[1,2] The distance coefficient was defined as follows:

$$d = \sqrt{\sum_{i=1}^{n} (X_i - Y_i)^2} \quad (2)$$

Where X*i* is the fragment size with index of i, and Y*i* is the average value of all samples. n is the total number of valid peaks. The value of d was integer more than zero.[3] The smaller value means the smaller variation among different replicates.

TABLES

Table S2. Calculation of Euclidean Distance Coefficient for Type 16 and Type 18

sample	Euclidean distance coefficient				
	type 16			type 18	
	Rsa Ⅰ	HaeⅢ	Pst Ⅰ	Rsa Ⅰ	Pst Ⅰ
1	5.85	1.33	2.52	2.68	0.47
2	8.45	1.33	5.77	4.28	0.75
3	4.70	3.67	9.55	6.87	3.40
4	1.69	4.33	5.17	8.59	4.96
5	7.56	5.67	4.54	8.12	0.47
6	2.64	4.33	2.00	3.28	4.68
7	8.48	3.33	2.34	4.28	2.36
8	6.94	1.67	6.15	5.32	0.75
9	3.38	3.67	1.89	3.46	4.68

2. HPV detection for clinical samples

2.1 Sequencing results for clinical samples

All PGMY-PCR products were addressed to T's and A's cloning and sequencing to confirm the RFLP-MCE results. The PCR products were separated on agarose gels

and target fragments were recovered from gels for cloning. White bacterial colonies indicated successful recombinant clone. Five positive colonies were selected to cultivate at 37℃ overnight. The harvested plasmids were added into PCR sequencing reaction and the final products were purified by method of NaAc /ethanol before loaded onto 3730 DNA analyzer.

Nine out of twelve positive infection samples (sample 1058, 2877, 2882, 2885, 3203, 3285, 3298, A3272, A3276) were successfully sequenced. The resulting sequences for each sample and their corresponding Genbank ID and type were shown in Table S4.

TABLES

Table S3

Table S3. Sequencing Blasting Results for Clinical Samples

sample code	HPV type	measured sequence	GenBank sequence
3203	6	TGCACAGGGACATAACAATGGTATTTGTTGG GGTAATCAACTGTTTGTTACTGTGGTAGATAC CACACGCAGTACCAACATGACATTATGTGCAT CCGTAACTACATCTTCCACATACACCAATTCT GATTATAAAGAGTACATGCGTCATGTGGAAG AGTATGATTTACAATTTATTTTTCAATTATGTA GCATTACATTGTCTGCTGAAGTAATGGCCTAT ATTCACACAATGAATCCCTCTGTTTTGGAAGA CTGGAACTTTGGGTTATCGCCTCCCCCAAATG GTACATTAGAAGATACCTATAGGTATGTGCA GTCACAGGCCATTACCTGTCAAAAGCCCACTC CTGAAAAGGAAAAGCCAGATCCCTATAAGAA CCTTAGTTTTTGGGAGGTTAATTTAAAAGAAA AGTTTTCTAGTGAATTGGATCAGTATCCTTTG GGACGA	HE962030
3298	16	TGCGCAGGGCCACAATAATGGCATTTGTTGG GGTAACCAACTATTTGTTACTGTTGTTGATAC	JF728174

		TACACGCAGTACAAATATGTCATTATGTGCTG	
		CCATATCTACTTCAGAAACTACATATAAAAAT	
		ACTAACTTTAAGGAGTACCTACGACATGGGG	
		AGGAATATGATTTACAGTTTATTTTTCAACTG	
		TGCAAAATAACCTTAACTGCAGACGTTATGAC	
		ATACATACATTCTATGAATTCCACTATTTTGG	
1058	18	TCGACCAAGGGGATATTGATCTAAGTCTAAA	JQ917454
		GAAAACTTTTCCTTTAAATCCACATTCCAAAA	
		CTTTAACTTATCATAGGGATCCTTATTTTCAG	
		CCGGTGCAGCATCCTTTAGACAGGTAATAGC	
		AACAGATTGTACAAAACGATATGTATCCACC	
		AAACTAGTAGTTGGCGGGGGGGGAACACCAA	
		AGTTCCAATCCTCTAAAATACTGCTATTCATA	
		CTATGAATATAGGACATAACATCTGCAGTTAA	
		AGTAATAGTACACAACTGAAAAATAAACTGC	
		AAATCATATTCCTCAACATGTCTGCTATACTG	
		CTTAAATTTGGTAGCATCATATTGCCCAGGTA	
		CAGGAGACTGTGTAGAAGCACATATTGTTAA	
		ATTGGTACTGCGAGTGGTATCTACCACAGTAA	
		CAAATAATTGATTATGCCAGCAAACACCATTG	
		TTATGTCCCTGTGCA	
2882	51	TGCGCAGGGCCACAATAATGGCATTTGCTGG	M62877
		AACAATCAGCTTTTTATTACCTGTGTTGATAC	
		TACCAGAAGTACAAATTTAACTATTAGCACTG	
		CCACTGCTGCGGTTTCCCCAACATTTACTCCA	
		AGTAACTTTAAGCAATATATTAGGCATGGGG	
		AAGAGTATGAATTGCAATTTATTTTTCAATTA	
		TGTAAAATTACTTTAACTACAGAGGTAATGGC	
		TTATTTACACACAATGGATCCTACCATTCTTG	
		AACAGTGGAATTTTGGATTAACATTACCTCCG	
		TCTGCTAGTTTGGAGGATGCATATAGGTTTGT	
		TAGAAATGCAGCTACTAGCTGTCAAAAGGAC	
		ACCCCTCCACAGGCTAAGCCAGATCCTTTGGC	
		CAAATATAAATTTTGGGATGTTGATTTAAAGG	
		AACGATTTTCTTTAGATTTAGACCAATTTGCA	
		TTGGGTCGCA	
3203	52	TCGACCTAAAGGAAACTGATCTAAATCTGCA	AB819274
A3276		GAAAACTTTTCTTTTAAATCCACCTCCCAAAA	
		CATATAGTCCTTTAAAGGATCTTCCTTTCCTTT	
		AGGTGGTGTGTTTTTTTGACAAGTTATAGCAG	
		TAGAAGTTACAAATCTGTATGTGTCCTCCAAA	
		GATGCAGACGGTGGTGGGGTAAGGCCAAATT	
		GCCAGTCCTCTAAAATAGTGGCATCCATCTTA	
		TGAATATATGTCATAACATCAGCTGTTAATGT	

		AATCTTGCACAATTGAAAAATAAATTGTAAAT CGAATTCCTCGCCATGACGAAGGTATTCCTTA AAATTTTCATTTTTATATGTGCTTTCCTTTTTC ACCTCAGCACATAAAGTCATGTTAGTGCTACG AGTGGTATCCACAACTGTGACAAACAACTGA TTGCCCCAACATATGCCATTATTGTGGCCCTG CGCA	
1058	53	TGCCAAGGGGAAACTGATCCAAATCAGCAGA AAAACTGTTTTGCAAATTGACCTCCCAAAATT TATATTTAGATAGTGGGTCCTGCTTTTCAGGA GGGGGCTGATCCTTTTGACAGGTTATAGCTGC ACTTTTCACATATCTGTATTTGTCCTCTAAGCT AGTGGCAACAGGAGGCGACAAACCTATATTC CAGTCTTCCAGTAAGGTAGAATTCATAGTATG TAAATAGGCCATAACCTCAGCAGACAGGGAT ATTTTACATAGTTGAAACACAAATTGTAATTC ATATTCCTCTGCATGTCTAACATACTGTTTAA TTTGCTTTGAATTATATGTAGACATAGACTGT GTGGTTGCGGAAAGAGTCATGTTTGTATTCCT GGTGGTATCCACAACAGTTACAAATAACTGA TTGTTCCAACAGATGCCATTATTATGTCCCTG TGCA	EF546475
3285 2885 2877 A3276	62	TCGACCCAAGGGAAACTGGTCCAAATCAGTA GACAACTTGTCCTTAAGATCCACAGTCCAAAA TGTCATTTGCGCATACGGGTCCACCTTGGGGG ACGGGGAAGCAGCCCCCTTTTGACATGTAAT AGCCCGAGACTGCAAATAGTGATATGTCTCAT CTAAACTAGTGGAAGGGGGGGTAAAACCCCA AAGTTCCAGTCATCCAAAAGGTCCTTGTTCAT ATTATGCAGGTAGGCCATGATTTCGGGGGTTA ACTGTATTTTGCACAATTGAAATATAAATTGC AAATCAAATTCCTCCGTGTGTCGCAAAAATTC CCTAAAGTTGGTAGCCTTGTATTCTGCTGCAG CAGTGGAGGCGGTACAAATAGTAAAATTAGT ACTCCTAGTAGTATCCACCACAGTAACAAAC AGTTCATTAAACCAACAAATACCATTATTGTG GCCCTGCGCA	AY395706
A3272	66	TCGTCCCAAAGGAAACTGATCTAGGTCTGCA GAAAAGCTGTCCTGTAAATTAACCTCCCAAA ACTTATATTTAGCCAGGGGATCCTGCTTTTCT GCAGGGGGCTGTTCCCTTTGACATGTAATAGC TGTGCTTTTAATATACCTATATTTATCCTCCAA GCTAGTTGCAACTGGTGGGGACAATCCAATG TTCCAATCGTCTAATAAAGTATTATTCATATT	DQ486474

		ATGCAAATATGCCATAACTTCTGCAGTTAAGG TTATTTTACAAAGTTGAAACACAAACTGTAGT TCATATTCCTCCACATGGCGAAGGTATTGATT GATTTCACGTGCATCATATTTAGTTAATGTGC TTTTAGCTGCATTAATAGTCATGTTGGTACTT CTGGTAGTATCCACAACAGTAACAAATACCT GATTACCCCAGCATATGCCATTGTTATGTCCC TGTGCA	
A3276	68	TCGTCCTAAAGGAAACTGGTCCAGTTCAGAA CTAAACTTTTCCTTTAAATTTACATTCCAAAA GTTTAAGCCATCATATGGATCCTTTTTAGTAG GTGCAGGGGCGTCTTTTTGACATGTAATTGCT GCTGATTGCAGATAGCGGTATGTATCTACAAG ACTAGCAGATGGTGGAGGGGCAACACCAAAA TTCCAATCATCCAAAATAGCAGGATTCATAGT ATGTATATATGACATTACATCAGTTGACAATG TTATAGAACACAACTGAAATATAAATTGCAA ATCATATTCCTCCACATGCCTAATATATTCCTT AAATTTATTAGGATCATAAATATTTGGTACAG CTGATTCAGTAGTAGTAGACAAAGTAAAATT GGTACTGCGAGTGGTATCCACAACAGTAAGA AATAATTGATTATGCCAACAAATACCATTGTT ATGTCCCTGTGC	JQ902131

References:

(1) Y. Chen, S.-B. Zhu, M.-Y. Xie, S.-P. Nie, W. Liu, C. Li, X.-F. Gong, Y.-X. Wang, *Anal. Chim. Acta*, 2008, **623**, 146-156.

(2) H. Lapid, S. Shushan, A. Plotkin, H. Voet, Y. Roth, T. Hummel, E. Schneidman, N. Sobel, *Nat. Neurosci.*, 2011, **14**, 1455-1461.

(3) A. B. Yongye, K. Byler, R. Santos, K. Martínez-Mayorga, G. M. Maggiora, J. L. Medina-Franco, *J. Chem. Inf. Model.*, 2011, **51**, 1259-1270.

2.2 Cytologic results for clinical samples

2.2.1 Preparation of Pap Smear for Cytologic Test

An aliquot of each kind of LBC sample were centrifuged at 1,500 rpm for 5 min. The cells sank down and adhered onto the surface of slide for 20 min. Pap smear was made according to the protocol of Papanicolaou test. In general, cells were fixed on slides with ethanol of a series of concentration. Nuclear staining was done with hematoxylin. Subsequencely, counterstains orange G and EA were used for cytoplasmic staining. The slides were cleared and mounted by cleaned water.

Cytologic test was performed for all clinical samples to screen out HPV infection samples. Main abnormal features were koilocytes indicating HPV infection and perinuclear halos indicating inflammation (Figure S2). However, most samples showed no significant cell abnormalities (Figure S3).

2.2.2 Cytologic results for samples with abnormalities

FIGURES

Figure S2

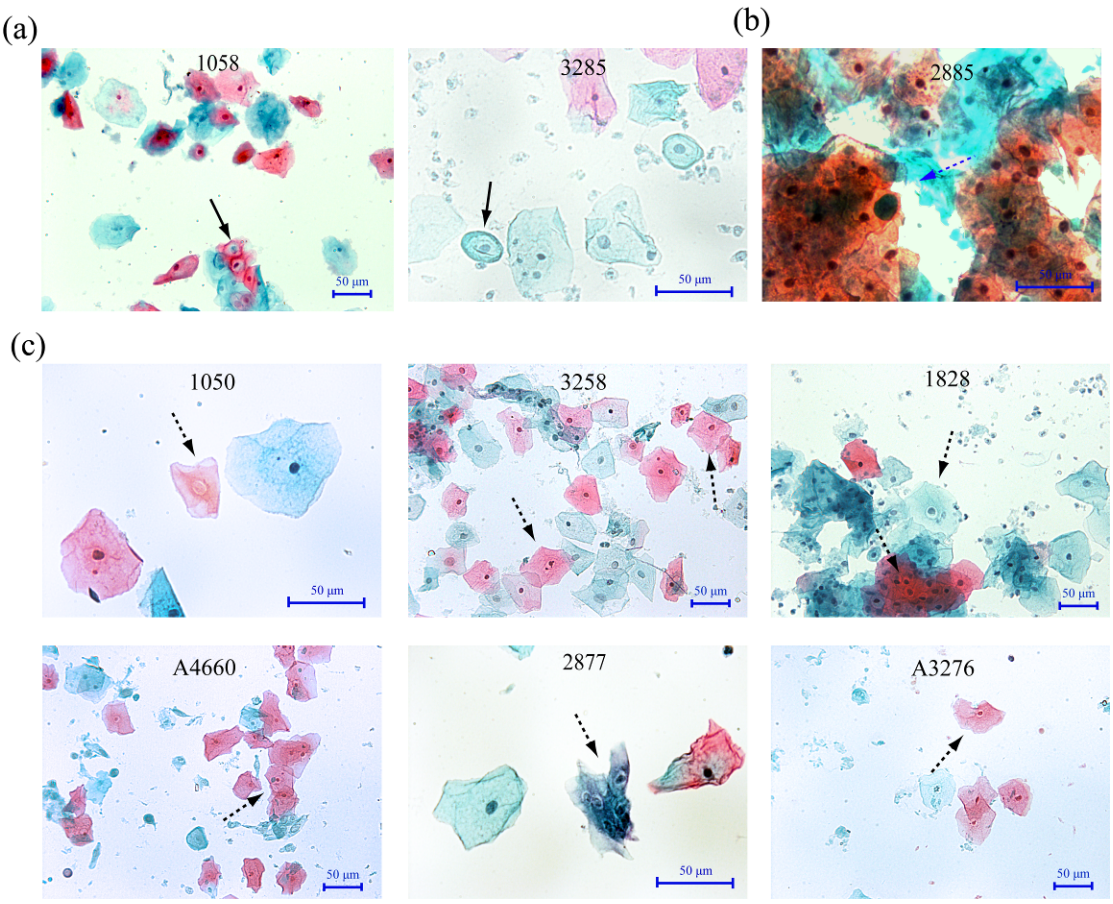

Figure S2. Cytologic micrographs were shown for sample 1058, 3285, 2885, 1050, 3258, 1828, A4660, 2887 and A3276. (a) The black arrow indicates the koilocytes. (b) The dotted blue arrow indicates the nuclear abnormality. (c) The dotted arrow in black indicates the perinuclear halos.

2.2.3 Cytologic results for samples without significant abnormalities

Figure S3

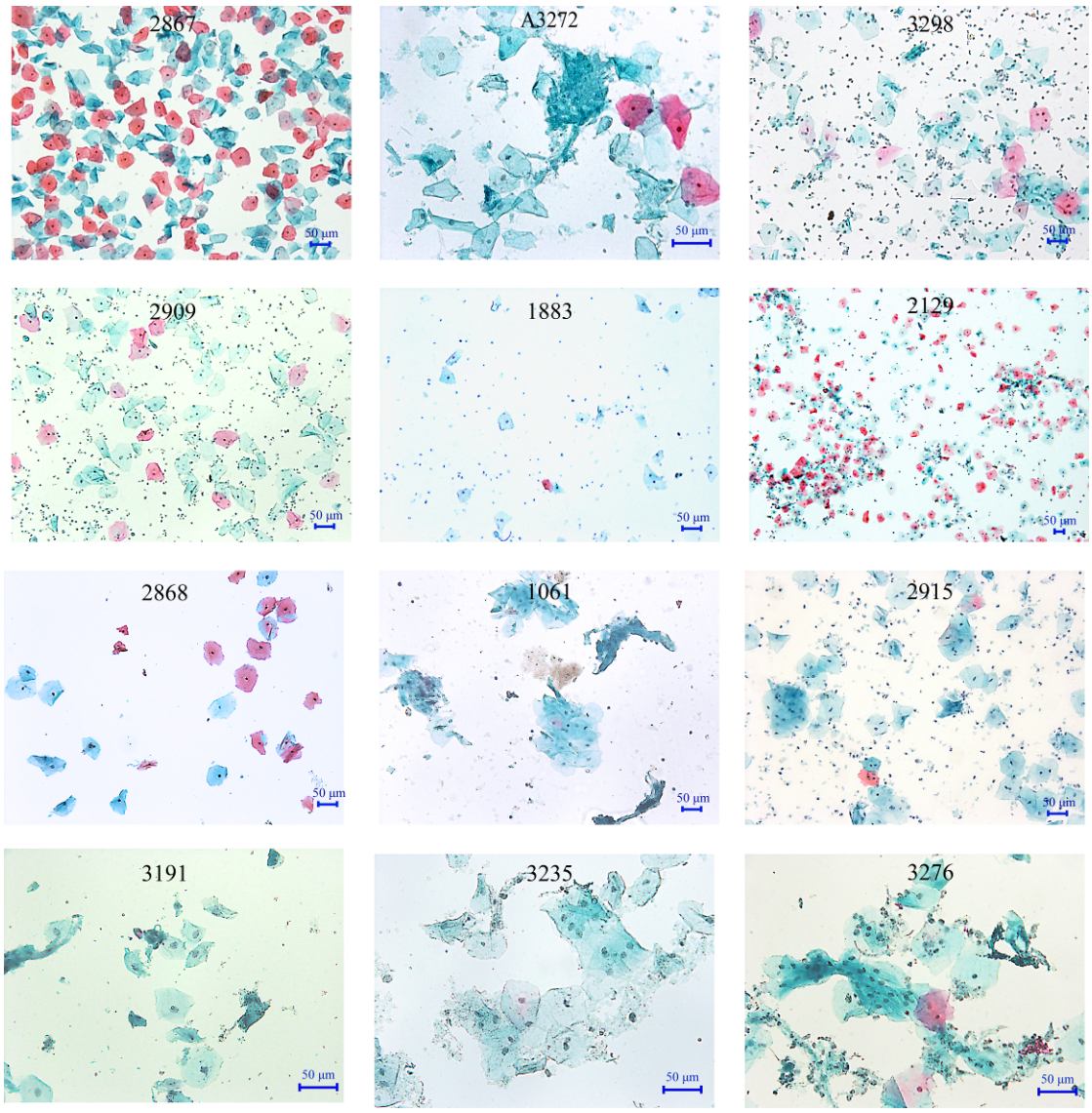

Figure S3. Cytologic micrographs were shown for sample 2867, A3272, 3298, 3909, 1883, 2129, 2868, 1061, 2915, 3191, 3235 and 3276. All these samples performed no significant cell abnormalities.

2.3 PCR-RFLP-MCE detection for clinical samples

PCR-RFLP-MCE typing method started with amplification of beta-globin gene and HPV L1 gene simultaneously. All DNA extracts were demonstrated to be adequate without influence of inhibitors (Figure S4). PCR amplicons of fragments about 450 base pairs in length for HPV L1 gene were detected in sample 1050, 1058, 2867,

2877, 2882, 2885, 3203,3258, 3285, 3298, A3272 and A3276 (Figure S5). And the corresponding identification maps for these samples were shown in Figure S6.

2.3.1 Beta-globin gene amplification

FIGURES

Figure S4

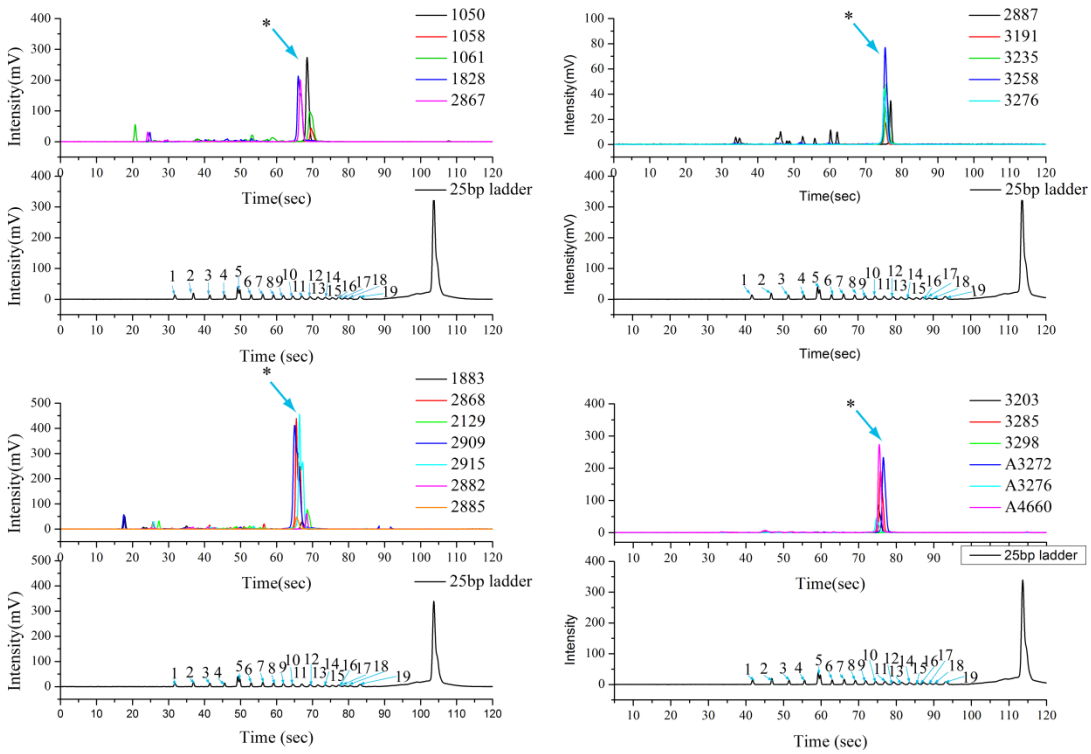

Figure S4. Beta-globin amplification were shown for clinical samples. Asterisk （*） indicated the target amplicons in the size of 268-bp. All samples can be observed the 268-bp fragment, proved to be valid for next amplification.

3.3.2 HPV gene amplification

Figure S5

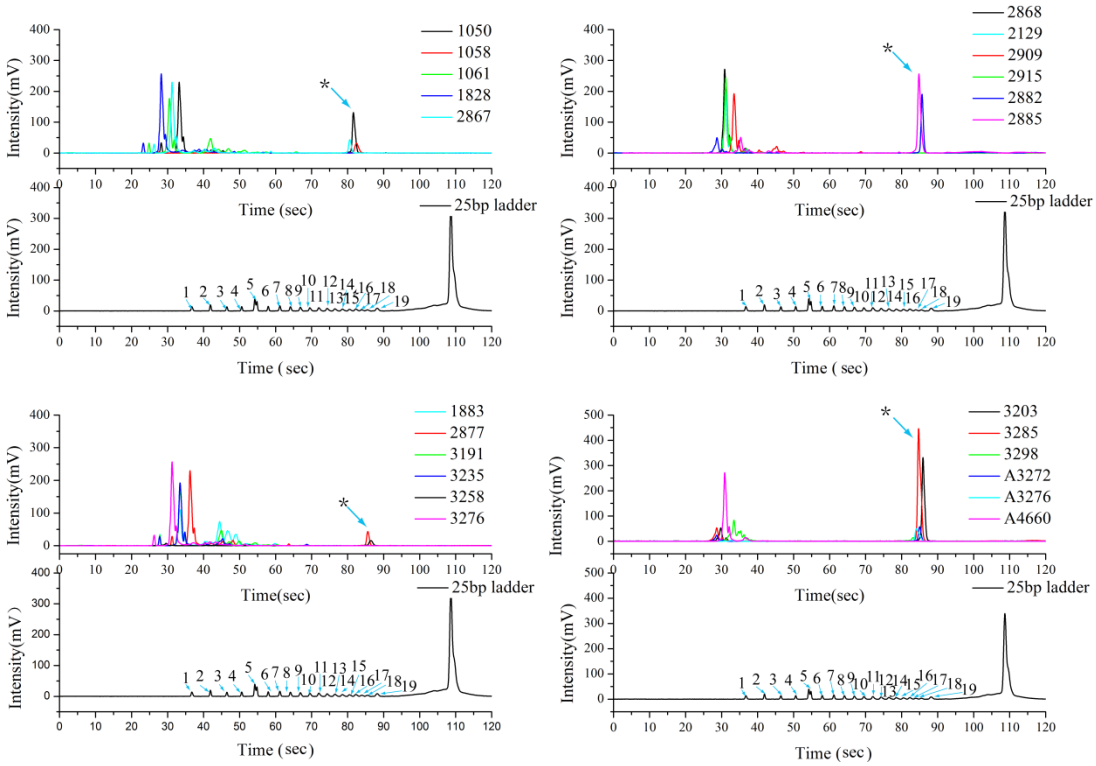

Figure S5. HPV infection screening results were shown for clinical samples. Asterisk （*） indicated the target amplicons in the size of about 450-bp.Samples indicated as 1050, 1058, 2867, 2877, 2882, 2885, 3203, 3258, 3285, 3298, A3272 and A3276 were HPV positive.

2.3.3 Identification maps for clinical samples

Figure S6

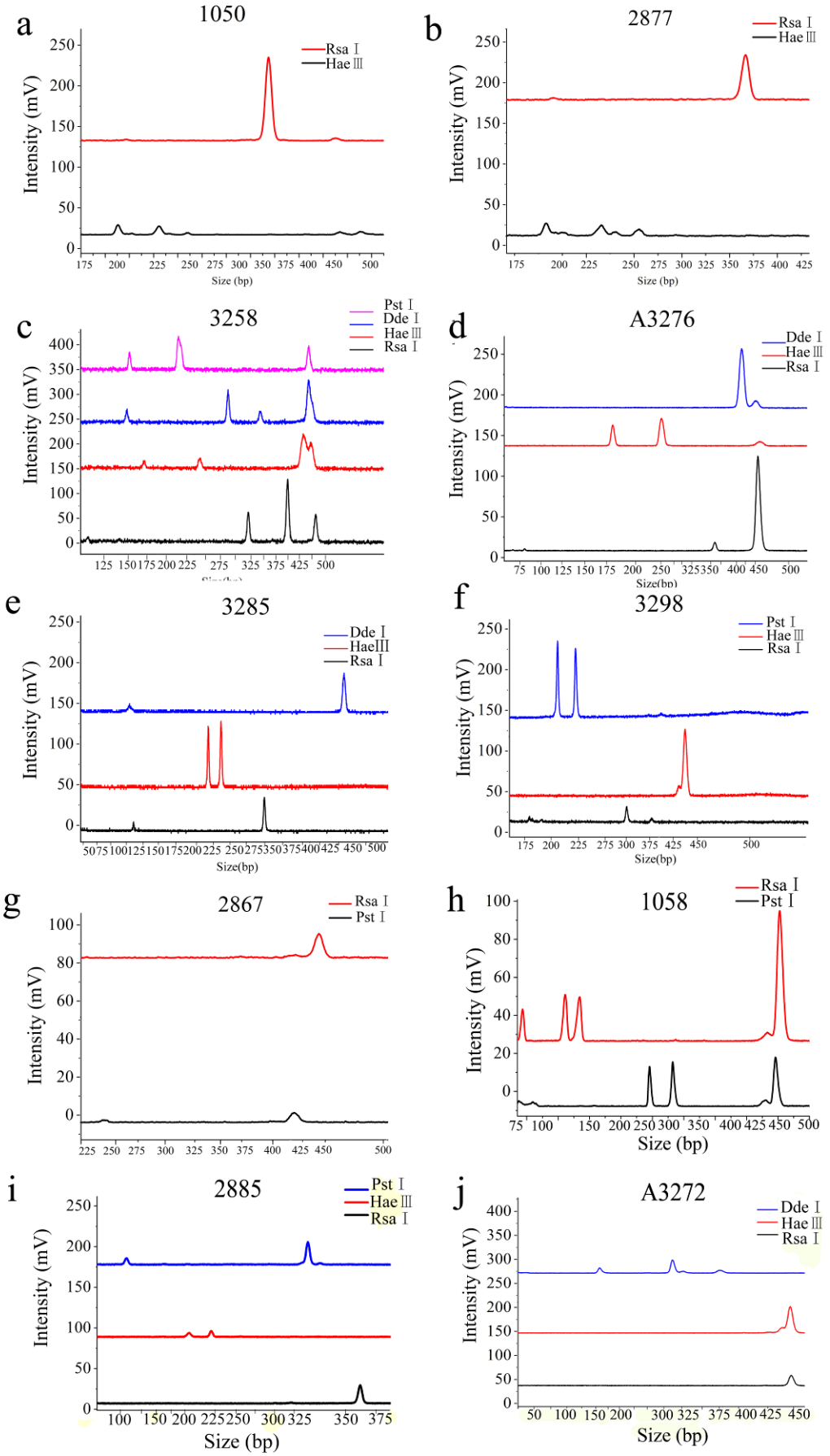
a
1050
Rsa I
Hae III
Intensity (mV)
Size (bp)
b
2877
Rsa I
Hae III
Intensity (mV)
Size (bp)
c
3258
Pst I
Dde I
Hae III
Rsa I
Intensity (mV)
d
A3276
Dde I
Hae III
Rsa I
Intensity (mV)
Size(bp)
e
3285
Dde I
HaeIII
Rsa I
Intensity (mV)
Size(bp)
f
3298
Pst I
Hae III
Rsa I
Intensity (mV)
Size(bp)
g
2867
Rsa I
Pst I
Intensity (mV)
h
1058
Rsa I
Pst I
Intensity (mV)
Size (bp)
i
2885
Pst I
Hae III
Rsa I
Intensity (mV)
Size (bp)
j
A3272
Dde I
Hae III
Rsa I
Intensity (mV)
Size (bp)

Figure S6. Results of identification maps were shown for samples (a) 1050, (b) 2877, (c) 3258, (d) A3276, (e) 3285, (f) 3298, (g) 2867, (h) 1058, (i) 2885 and (j) A3272. HPV types can be identified from the maps as HPV 72 for sample 1050, HPV 72 for sample 2877, HPV 62 and 68 for A3276, HPV 16 for sample 3298, HPV 62 for sample 3285 and 2885, HPV 52 or 59 for sample 2867, HPV 18 and 53 for sample 1058, HPV 66 for sample A3272. The restriction pattern of 3258 was too complicated to match with certain types.

3. computational genotyping methods

3.1 The design of typing software

Calculation of the compatibility degree was based on the parameters of cosine of angle. The cosine of angle (cos α) can be calculated as follow[1]:

$$S = \cos\alpha = \sum_{i=1}^{n} X_i \times Y_i \div \left(\sqrt{\sum_{i=1}^{n} X_i^2} \times \sqrt{\sum_{i=1}^{n} Y_i^2} \right)$$

(1)

The value range of S is between 0 and 1, cos α =0 means no similarity, while cos α =1 means no difference.

Here, we assumed that at most four valid fragments would be produced after one kind of restriction endonuclease was used. Therefore, we defined dR_PD1, dR_PD2, dR_PD3and dR_PD4 as four collections corresponding to theoretical size sets produced by four enzymes. And we also defined nSxy as the cos α，nSx as X*i* and nSy as Y*i* .For each restriction endonuclease, the number of potential cleaved

fragments varies from 1 to 4. This has no effect on the calculation of cos α, and the program runs as follow:

```
nSxy=dR_PD1*aRsa(i,1)+dR_PD2*aRsa(i,2)+dR_PD3*aRsa(i,3)+dR_PD4*aRsa(i,
4)
  nSx=SQRT(dR_PD1^2+dR_PD2^2+dR_PD3^2+dR_PD4^2)
  nSy=SQRT(aRsa(i,1)^2+aRsa(i,2)^2+aRsa(i,3)^2+aRsa(i,4)^2)
  nSxSy=nSx*nSy
  nS=nSxy/nSxSy
nS1=round(nS,6)
```

The typing results read by typing software was finally shown as below:

FIGURES

Figure S7

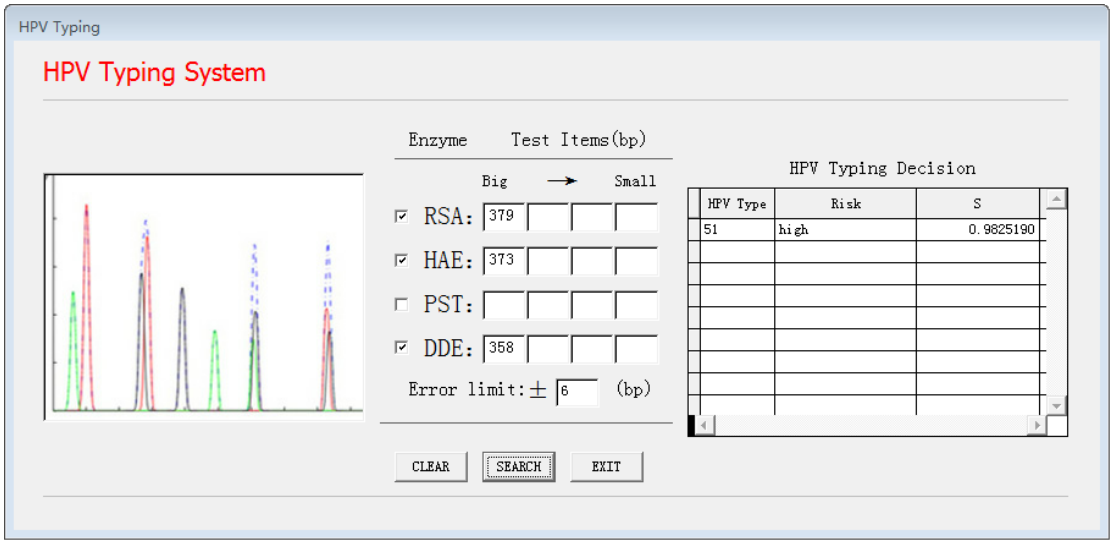

Figure S7. Example of typing results by typing software.

References:

(1) L. W. Yang, D. H. Wu, X. Tang, W. Peng, X. R. Wang, Y. Ma, W. W. Su, *J. Chromatogr. A* 2005, **1070**, 35-42.

3.2 Evaluation of typing results by compatibility degree

We verified the accuracy of the PCR-RFLP-MCE automatic typing method in 23 clinical samples, and calculated the cosine of angle (cos α) as compatibility degree obtained from positive samples expect sample 3258 in these specimens. All tested samples carried out good results with high compatibility degree of 94.21-100%.

TABLES

Table S4. Calculation of Compatibility Degree for 11 Real Samples

sample code	HPV type	a compatibility degree %
1050	72	98.11
1058	18	94.21
	53	100.00
2867	52	100.00
	59	100.00
2877	62	98.05
2882	51	98.25
2885	62	98.05
	69	98.11
3203	6	99.82
	52	100.00
3285	62	98.05