Infection, Genetics and Evolution 8 (2008) 689–692

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Short communication

Detecting recombination and diversifying selection in human alpha-papillomavirus

A. Carvajal-Rodríguez *

Dpto. de Bioquímica, Genética e Inmunología, Facultad de Biología, Universidad de Vigo, 36310 Vigo, Spain

ARTICLE INFO

Article history:
Received 21 May 2008
Received in revised form 4 July 2008
Accepted 8 July 2008
Available online 15 July 2008

Keywords:
HPV
Recombination
Positive selection

ABSTRACT

Intragenic recombination and selection analyses were performed in DNA sequences of human alpha-papillomavirus. Recombination was estimated and the corresponding breakpoints obtained by re-analyzing data grouped by phylogenetic and epidemiological criteria, using different alignment methods. Diversifying or positive selection has been scarcely studied in these viruses probably due to the high divergence rates. We have applied maximum likelihood, empirical Bayesian and maximum parsimony methods to detect the presence of positive selection. Within the HPV 16 type, significant positive selection was detected at the time of the separation of the African 1 and African 2 branches from the other populations. At the inter-type level, positive selection can be traced in some codons of the gene L2 of the high and low risk groups. These results indicate that positive selection could have been important in the evolution of HPV both at inter- and intra-type levels.

1. Introduction

Human papillomavirus (HPV) is etiological agent of cervical cancer with carcinogenic HPV causing virtually all cervical cancers (Monsonego et al., 2004). Phylogenetic studies of papillomavirus have shown that different genes have different evolutionary rates between and within PV types (Garcia-Vallve et al., 2005; Gottschling et al., 2007). Hence, papillomavirus diversification has been driven by multiple mechanisms, including recombination, and not only by co-evolution with the corresponding hosts (Gottschling et al., 2007). In a previous work (Angulo and Carvajal-Rodriguez, 2007), recombination was estimated in several human alpha-papillomavirus grouped by phylogenetic and epidemiological criteria and the first evidence of HPV intra-type recombination was provided. In the present work, we have studied the HPV genes L1 and L2, coding for structural proteins, and genes E6 and E7 that regulate host-cell DNA replication and transformation. We have confirmed most of previous recombination signals by using improved alignment methods. By other way, diversifying selection has been scarcely studied in PV sequences (Halpern, 2000). When high level of divergence exists, saturation of genetic changes impedes satisfactory selection analysis. However, closely related types have allowed selection analysis prior to saturation especially in the overlapping genes E2 and E4 (Narechania et al., 2005). For genes L1, L2, E6 and E7, no previous evidence of positive selection has been found at inter-type level though selection has been detected in E5 and E6 genes within HPV 16 type (Chen et al., 2005; DeFilippis et al., 2002).

As stated above, the study of diversifying selection at HPV inter-type level is difficult. When the species evolutionary distance is too large, the accuracy of the likelihood estimates decrease (Mayrose et al., 2004; Nei, 2005). Indeed, the maximum likelihood (ML) methods have been shown to be more affected by highly diverged sequences than the empirical Bayesian (EB) ones (Mayrose et al., 2004). Additionally, ML and EB are known to produce many false-positive results while maximum parsimony (MP) methods do not (Suzuki and Nei, 2004). On the other side, MP methods require a larger number of sequences for efficient detection of positive selection. Due to the low percentage of sequence identity in the different HPV data sets we performed various estimation strategies, including simulations, to try to get reliable information on synonymous and non-synonymous substitution rates.

2. Results and discussion

2.1. Recombination

After processing DNA sequences with the program Gblocks (Castresana, 2000), the recombination signal in L1GI and E6GII data sets was detected in less than 50% of the estimates and therefore

* Tel.: +34 986 813828; fax: +34 986 812556.
E-mail address: acraaj@uvigo.es.

1567-1348/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.meegid.2008.07.002

Table 1
Recombination breakpoints in the α-HPV sequences

	HPV16	GI	GII	GIII
E6	– (16)	214 (313)	– (331)	– (294)
E7	171 (6)	– (189)	– (168)	– (149)
L1	– (38)	– (857)	– (879)	– (603)
L2	– (61)	226, 437, 912, 1035 (596)	293, 577, 1030 (736)	166, 1029 (714)

A dash implies no significant recombination or no break point detected. The total number of segregating sites appears in parenthesis. Groups are the same as in (Angulo and Carvajal-Rodriguez, 2007). Group I (GI) included the 14 most common high-risk types (16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 73, 82; n = 14 sequences including just one variant of type 16). Group II (GII) included 6 low risk types (6, 11, 40, 42, 43, 44; n = 8 sequences including the 3 variants of type 6). Group III (GIII) included 3 low risk types plus 5 undetermined risk types which cluster together (61, 72, 81, 62, 71, 83, 84, 89; n = 12 or 11 sequences including 5 or 4 variants of type 71. For this group, L1 has 13 sequences because of 4 variants from type 71 plus 2 additional variants from types 72 and 81). HPV16: The accession numbers for this group are provided in Fig. 1.

considered as non-significant. Additionally, an intra-type recombination break point was detected at nucleotide 171 (codon 57) in the HPV16 E7 gene (Table 1). Therefore, almost all recombination signals found in a previous study (Angulo and Carvajal-Rodriguez, 2007) were confirmed after elimination of saturated and non-homologous positions and using improved alignment methods. The presence of recombination both at intra and inter-type levels is interesting since there is growing evidence of multiple infection with different HPV types (Antonishyn et al., 2008) increasing the possibility of current HPV recombination.

2.2. Positive selection

After the recombination analysis, we studied the presence of diversifying selection in the same HPV DNA data sets. We have performed maximum likelihood (ML), maximum parsimony (MP) and empirical Bayesian (EB) analyses (Table 2). Within the HPV16 group, both ML and EB methods detected codons 10, 14 and 83 in gene E6 as positively selected ($dN/dS > 1$). However, only under EB they were statistically significant. Diversifying selection was earlier detected in these same codons by previous studies (Chen et al., 2005; DeFilippis et al., 2002) using the ML method. The polymorphism 83L/V has been significantly associated with persistent infection and therefore with tumor development (Lee et al., 2008). E6 is responsible for direct interaction with the cellular protein p53 that could cause the selective pressure onto the HPV gene due to the differential ability to interact with the different p53 variants (DeFilippis et al., 2002). In addition, there is evidence of varying degrees of association of the HPV16 variants with cervical cancer, seeming that the non-European variants (African and Asian-American) have a higher risk of provoking it (Schiffman et al., 2005; Sichero et al., 2007). Consistently with our positive selection analysis, a significant pattern was found only in the gene E6 when studying the variation in selection pressure through time within the HPV16 group. As can be appreciated in Fig. 1, the presence of positive selection appears through the different branches of the tree. Importantly, there is high confidence (>99%) in the effect of diversifying selection at the time of the separation of the African 1 and African 2 branches from the other populations. These findings are interesting since differences between African and non-African populations have been reported regarding the E6 target gene p53 (Katkoori et al., 2004; Porter et al., 2004). Therefore, it could be that coevolution with cell-cycle regulators has been driven the divergence in the E6 genes. Furthermore, it seems that the above-mentioned codons suffice to classify the HPV16 variants. For example, codon 10 separate African 1 and 2 while codon 14 distinguishes African 2 from the non-African variants. Asian-American populations are characterized by the 10R14H83V combination while codon 83 seems to have undergone selection indeed within the European populations.

When studying inter-type groups and due to the low percentage of sequence identity in the different HPV data sets (Table 2), various estimation strategies were performed. For instance, the fixed effects likelihood (FEL) method detects more sites than any other, probably due to larger type 1 error (false positives). In fact, false positive explosion occurs in some of the data sets both under the ML and EB methods. The FEL method is a conservative test and therefore its false positive rate is expected to be very low (Kosakovsky Pond and Frost, 2005a). However, the rate could increase with the sequences divergence (Mayrose et al., 2004; Suzuki and Nei, 2004). To check this point, we carried out simulations (see Methods and Supplementary material on line, for details) in order to estimate the frequency of false positives due to low percentage (60–70%) of sequence identity. As a result, we have confirmed the tendency of false positive explosion when using the FEL method with high divergent sequences. In 100% of the replicates, about 8% of the sites were spuriously estimated as positively selected. One possibility to deal with this problem is to apply multi-test correction. Nevertheless, this is not an efficient strategy because the null hypothesis is not the same for every site (due to purifying selection) and because the key question is whether any sites are under selection rather than identifying particular sites (Massingham and Goldman, 2005). An alternative approach is to use simulations to estimate the highest significance level to avoid false positive explosion given the corresponding evolutionary settings. Thus, we performed simulations using the parameters estimated from the real data and check that the maximum P value allowing for no false positive detection was 0.5%. Therefore, under the FEL method, we set this significance level to assign a codon as positively selected (Table 2). Interestingly 63% (5/8) of the sites detected in this way coincide with those detected by the most conservative MP method. With the MP method, 100% of the sites were also detected using some other approach, as expected due to the low power of the MP methodology. Only codon

Table 2
Number of positively selected codons in the α-HPV sequences detected by ML, MP and EB methods

	ML10%	ML0.5%	MP	EB	Positively selected codons
E6HPV16 (99)	1*	0	0	3	10^{4}, 14^{4}, $83^{1,4}$
E7HPV16 (99)	0	0	0	0	
L1HPV16 (99)	0	0	0	0	
L2HPV16 (99)	0	0	0	1	378^{4}
E6GI (64)	1	0	0	0	
E7GI (62)	2	0	0	0	
L1GI (71)	0	0	0	0	
L2GI (63)	64	5	4	1	$111^{2,3}$,$190^{2,3}$, $293^{2,3}$, 327^{2}, $346^{2,3}$, 466^{4}
E6GII (62)	1	0	0	0	
E7GII (65)	0	0	0	0	
L1GII (71)	5	0	0	0	
L2GII (66)	47	1	1	41	$64^{2,3,4}$
E6GIII (65)	2	0	0	1	129^{4}
E7GIII (72)	4	0	0	4	16^{4}, 67^{4}, $83^{1,4}$, 93^{4}
L1GIII (75)	81	2	1	56	$387^{3,4}$
L2GIII (72)	5	0	0	0	

The % of sequence identity is shown in parenthesis. MLP%: Maximum likelihood with P%/100 significance level. MP: maximum parsimony EB: empirical Bayesian. *Selection at internal branches of the tree. Superscripts at positively selected codons identify the detection method by its order in the table. Codons corresponding to positives with ML10% (superscript 1) method were only reported when detected by some other method. Note that any codon detected by ML0.5% (superscript 2) is necessarily detected by ML10%. When just detected by FEL (ML10% and ML0.5%) only codons with dN and $dS > 0$ are reported in the last column. The codon positions in the last column refer to the original sequences with gaps and stop codons.

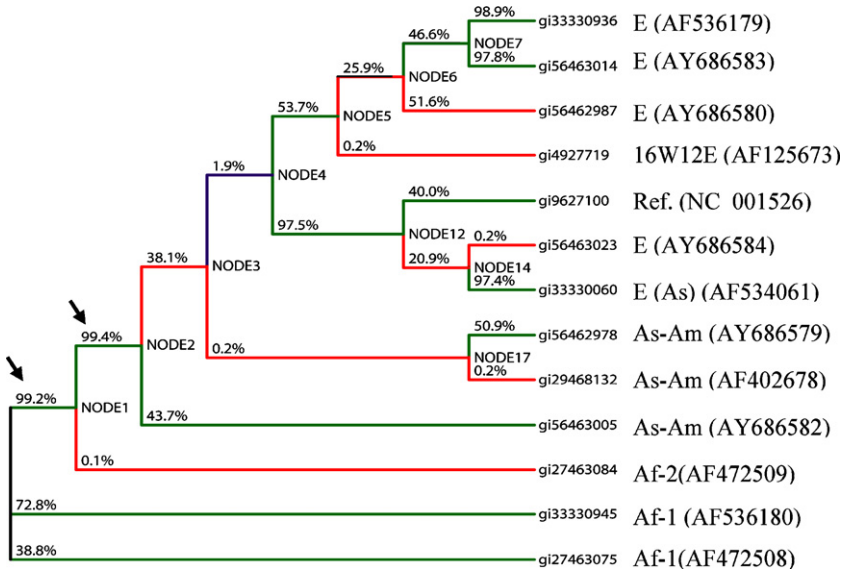

Fig. 1. Variation in selection pressure through time in gene E6 within the HPV16 group. Branch labels represent model averaged probabilities of d*N*/d*S* > 1. The composition of the distinct variants for the three positively selected codons (10, 14 and 83) is given.

64 in the L2GII data set was detected by all the methods. As a conclusion, though it is true that the detection of positive selection between the highly divergent HPV types should be considered with caution, it seems that diversifying selection can be traced in the gene L2 of the high (GI) and low (GII) risk groups. In the undetermined-risk group (GIII), at least, codons 83 and 387 in genes E7 and L1 respectively seem to have undergone diversifying selection, as well.

3. Methods

HPV DNA sequences were grouped attending to phylogenetic and epidemiological information, as well as to clinical manifestations. Sequence accession numbers are the same as in a previous work (Angulo and Carvajal-Rodriguez, 2007). In the HPV16 group, the haplotypes used are representative of the major HPV16 variants (Chen et al., 2005). The accession numbers for this group are provided in Fig. 1.

3.1. Sequence alignment

For each sequence group, sequences were aligned with ClustalX (Thompson et al., 1997). In addition, new alignment was performed using the FFT-NS-2 and INS-i algorithms as implemented in the Mafft program (Katoh et al., 2005). To eliminate saturated or non-homologous positions the alignments were processed with GBlocks (Castresana, 2000). Stop codons and gaps were eliminated. Whatever the alignment method, the best-fit model of nucleotide substitution was selected under the Akaike information criteria with Modeltest v3.6 (Posada and Crandall, 1998), using maximum likelihood estimates from PAUP* (Swofford, 2002). Maximum likelihood trees were estimated under the best-fit model using the algorithm implemented in Phyml v.2.4.1 (Guindon and Gascuel, 2003). Maximum parsimony trees (Hartigan, 1973) were obtained using the program NJboot (Takezaki et al., 1995).

3.2. Recombination estimation and breakpoint detection

The composite likelihood estimator and its permutation test (McVean et al., 2002) as implemented in the *Kpairwise* program (Carvajal-Rodriguez et al., 2006) which allows for complex nucleotide models and rate variation among sites were used to study population recombination rate. Five independent estimates were performed for each data set under a gene conversion model. When significant recombination signal was detected the genetic algorithm recombination detection (GARD) method (Kosakovsky Pond et al., 2006) was used to identify the recombination breakpoints.

3.3. Selection detection

Maximum likelihood (ML), maximum parsimony (MP) and empirical Bayesian (EB) methods were used to estimate selection. Under the ML framework the fixed effects likelihood model (FEL) (Kosakovsky Pond and Frost, 2005a) that estimates d*N*/d*S* on a site-by-site basis was used to detect selection (Kosakovsky Pond et al., 2005). If recombination was detected in a given data set different tree partitions were used to perform the selection analysis (Kosakovsky Pond and Frost, 2005c). The GA-branch approach (Kosakovsky Pond and Frost, 2005b) as implemented in Datamonkey server (Kosakovsky Pond and Frost, 2005c) uses a genetic algorithm-based procedure to detect different selective regimes through the branches of a tree. The MP analysis was performed using the programs NJboot and Adaptsite (Suzuki et al., 2001; Takezaki et al., 1995). The EB approach was performed with the program Selecton 2.2 (Stern et al., 2007) using its highest precision level and the ML trees previously estimated with the program Phyml. Additionally, two simulations of 100 datasets with 333 codons each were performed under neutral conditions to get sequences with 60%, 70% and 98% of identity using the software Genomepop (Carvajal-Rodriguez, 2008). Details on the simulations are given in the Supplementary material.

Acknowledgement

AC-R is currently funded by an Isidro Parga Pondal research fellowship from Xunta de Galicia (Spain).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.meegid.2008.07.002.

References

Angulo, M., Carvajal-Rodriguez, A., 2007. Evidence of recombination within human alpha-papillomavirus. Virol. J. 4, 33.

Antonishyn, N.A., Horsman, G.B., Kelln, R.A., Saggar, J., Severini, A., 2008. The impact of the distribution of human papillomavirus types and associated high-risk lesions in a colposcopy population for monitoring vaccine efficacy. Arch. Pathol. Lab. Med. 132, 54–60.

Carvajal-Rodriguez, A., 2008. GENOMEPOP: a program to simulate genomes in populations. BMC Bioinformatics 9, 223.

Carvajal-Rodriguez, A., Crandall, K.A., Posada, D., 2006. Recombination estimation under complex evolutionary models with the coalescent composite likelihood method. Mol. Biol. Evol. 23, 817–827.

Castresana, J., 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. 17, 540–552.

Chen, Z., Terai, M., Fu, L., Herrero, R., DeSalle, R., Burk, R.D., 2005. Diversifying selection in human papillomavirus type 16 lineages based on complete genome analyses. J. Virol. 79, 7014–7023.

DeFilippis, V.R., Ayala, F.J., Villarreal, L.P., 2002. Evidence of diversifying selection in human papillomavirus type 16 E6 but not E7 oncogenes. J. Mol. Evol. 55, 491–499.

Garcia-Vallve, S., Alonso, A., Bravo, I.G., 2005. Papillomaviruses: different genes have different histories. Trends Microbiol. 13, 514–521.

Gottschling, M., Stamatakis, A., Nindl, I., Stockfleth, E., Alonso, A., Bravo, I.G., 2007. Multiple evolutionary mechanisms drive papillomavirus diversification. Mol. Biol. Evol. 24, 1242–1258.

Guindon, S., Gascuel, O., 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52, 696–704.

Halpern, A.L., 2000. Comparison of papillomavirus and immunodeficiency virus evolutionary patterns in the context of a papillomavirus vaccine. J. Clin. Virol. 19, 43–56.

Hartigan, J.A., 1973. Minimum mutation fits to a given tree. Biometrics 29, 53–65.

Katkoori, V.R., Gandham, M., Jhala, N.C., Soong, R., Diasio, R.B., Meleth, S., Grizzle, W.E., Manne, U., 2004. Differences in the mutational spectra of the p53 gene in proximal colonic adenocarcinomas of African-Americans and Caucasians. AACR Meeting Abstracts, 930-.

Katoh, K., Kuma, K., Toh, H., Miyata, T., 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 33, 511–518.

Kosakovsky Pond, S.L., Frost, S.D., 2005a. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol. Biol. Evol. 22, 1208–1222.

Kosakovsky Pond, S.L., Frost, S.D.W., 2005b. A genetic algorithm approach to detecting lineage-specific variation in selection pressure. Mol. Biol. Evol. 22, 478–485.

Kosakovsky Pond, S.L., Frost, S.D.W., 2005c. Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. Bioinformatics 21, 2531–2533.

Kosakovsky Pond, S.L., Frost, S.D.W., Muse, S.V., 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics 21, 676–679.

Kosakovsky Pond, S.L., Posada, D., Gravenor, M.B., Woelk, C.H., Frost, S.D., 2006. Automated phylogenetic detection of recombination using a genetic algorithm. Mol. Biol. Evol. 23, 1891–1901.

Lee, K., Magalhaes, I., Clavel, C., Briolat, J., Birembaut, P., Tommasino, M., Zehbe, I., 2008. Human papillomavirus 16 E6, L1, L2 and E2 gene variants in cervical lesion progression. Virus Res. 131, 106–110.

Massingham, T., Goldman, N., 2005. Detecting amino acid sites under positive selection and purifying selection. Genetics 169, 1753–1762.

Mayrose, I., Graur, D., Ben-Tal, N., Pupko, T., 2004. Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. Mol. Biol. Evol. 21, 1781–1791.

McVean, G.A.T., Awadalla, P., Fearnhead, P., 2002. A coalescent based-method for detecting and estimating recombination from gene sequences. Genetics 160, 1231–1241.

Monsonego, J., Bosch, F.X., Coursaget, P., Cox, J.T., Franco, E., Frazer, I., Sankaranarayanan, R., Schiller, J., Singer, A., Wright Jr., T.C., Kinney, W., Meijer, C.J., Linder, J., McGoogan, E., Meijer, C., 2004. Cervical cancer control, priorities and new directions. Int. J. Cancer 108, 329–333.

Narechania, A., Terai, M., Burk, R.D., 2005. Overlapping reading frames in closely related human papillomaviruses result in modular rates of selection within E2. J. Gen. Virol. 86, 1307–1313.

Nei, M., 2005. Selectionism and neutralism in molecular evolution. Mol. Biol. Evol. 22, 2318–2342.

Porter, P., Lund, M., Lin, M., Yuan, X., Liff, J., Flagg, E., Coates, R., Eley, J., 2004. Racial differences in the expression of cell cycle-regulatory proteins in breast carcinoma. Cancer 100, 2533–2542.

Posada, D., Crandall, K.A., 1998. Modeltest: testing the model of DNA substitution. Bioinformatics 14, 817–818.

Schiffman, M., Herrero, R., Desalle, R., Hildesheim, A., Wacholder, S., Rodriguez, A.C., Bratti, M.C., Sherman, M.E., Morales, J., Guillen, D., Alfaro, M., Hutchinson, M., Wright, T.C., Solomon, D., Chen, Z., Schussler, J., Castle, P.E., Burk, R.D., 2005. The carcinogenicity of human papillomavirus types reflects viral evolution. Virology 337, 76–84.

Sichero, L., Ferreira, S., Trottier, H., Duarte-Franco, E., Ferenczy, A., Franco, E., Villa, L., 2007. High grade cervical lesions are caused preferentially by non-European variants of HPVs 16 and 18. Int. J. Cancer 120, 1763–1768.

Stern, A., Doron-Faigenboim, A., Erez, E., Martz, E., Bacharach, E., Pupko, T., 2007. Selecton 2007: advanced models for detecting positive and purifying selection using a Bayesian inference approach. Nucleic Acids Res. 35, W506–W511.

Suzuki, Y., Nei, M., 2004. False-positive selection identified by ML-based methods: examples from the Sig1 gene of the diatom Thalassiosira weissflogii and the tax gene of a human T-cell lymphotropic virus. Mol. Biol. Evol. 21, 914–921.

Suzuki, Y., Gojobori, T., Nei, M., 2001. ADAPTSITE: detecting natural selection at single amino acid sites. Bioinformatics 17, 660–661.

Swofford, D.L., 2002. PAUP* Phylogenetic Analysis Using Parsimony (*and Other Methods), 4 ed. Sinauer Associates, Sunderland, Massachusetts.

Takezaki, N., Rzhetsky, A., Nei, M., 1995. Phylogenetic test of the molecular clock and linearized trees. Mol. Biol. Evol. 12, 823–833.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., Higgins, D.G., 1997. The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 24, 4876–4882.